Overview of the CLEF eHealth Evaluation Lab 2019

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

[1] Maynooth University, Kildare, Ireland
liadh.kelly@mu.ie

[2] The Australian National University,
Data61/Commonwealth Scientific and Industrial Research Organisation,
University of Canberra, Canberra, ACT, Australia
hanna.suominen@anu.edu.au

[3] University of Turku, Turku, Finland

[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
Lorraine.Goeuriot@imag.fr

[5] German Centre for the Protection of Laboratory Animals (Bf3R),
German Federal Institute for Risk Assessment (BfR), Berlin, Germany
mariana.lara-neves@bfr.bund.de

[6] Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
{E.Kanoulas,D.Li}@uva.nl

[7] Computer and Information Sciences, University of Strathclyde, Glasgow, UK
leif.azzopardi@strath.ac.uk

[8] Cochrane Netherlands and UMC Utrecht,
Julius Center for Health Sciences and Primary Care, Utrecht, Netherlands
R.Spijker-2@umcutrecht.nl

[9] University of Queensland, Brisbane, Australia
{g.zuccon,h.scells}@uq.edu.au

[10] Qatar Computing Research Institute (QCRI), HBKU, Doha, Qatar
jpalotti@hbku.edu.qa

Abstract. In this paper, we provide an overview of the seventh annual edition of the CLEF eHealth evaluation lab. CLEF eHealth 2019 continues our 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 electronic health information in a multilingual setting. This year's lab advertised three tasks: Task 1 on indexing non-technical summaries of German animal experiments with International Classification of Diseases, Version 10 codes; Task 2 on technology assisted reviews in empirical medicine building on 2017 and 2018 tasks in English; and Task 3 on consumer health search in mono- and multilingual settings that builds on the 2013–18 Information Retrieval tasks. In total nine teams took part in these tasks (six in Task 1 and three

LK, HSu & LG co-chaired the lab. MN; EK, DL, LA & RS; and GZ, HSc & JP led Tasks 1–3, respectively.

F. Crestani et al. (Eds.): CLEF 2019, LNCS 11696, pp. 322–339, 2019.
https://doi.org/10.1007/978-3-030-28577-7_26

in Task 2). Herein, we describe the resources created for these tasks and evaluation methodology adopted. We also provide a brief summary of participants of this year's challenges and results obtained. As in previous years, the organizers have made data and tools associated with the lab tasks available for future research and development.

Keywords: Evaluation · Entity linking · Information retrieval · Health records · High recall · Information extraction · Medical informatics · Self-diagnosis · Systematic reviews · Test-set generation · Text classification · Text segmentation

1 Introduction

Retrieving, digesting, and summarising valid and relevant information to make health-centered decisions has become increasingly difficult in today's information overloaded society. More and more *electronic health* (eHealth) content is becoming available in a variety of forms ranging from scientific papers and health-related websites through patient records and medical dossiers to medical-related topics shared across social networks [27]. Laypeople, clinicians, and policy makers need bespoke systems to retrieve relevant and reliable contents and access them in a clear and concise way to easily judge and make sense of them to support their decision making.

Information retrieval (IR) systems have been commonly used as a means to access health information available online. To illustrate the immense worldwide popularity of going online to consume and produce health information, five years ago, in Australia, 40 per cent of searches were to fulfill health information needs; in Europe, nearly half of the population consider the Internet as a significant source of health information; and in the USA, nearly 70 per cent of people using web search engines want information about diseases, health conditions, or other medical disorders [1]. Based on the "Household Use of Information Technology" survey for 2016–2017 by the *Australian Bureau of Statistics* (ABS)[1], this popularity has grown and stabilised itself to almost 90 per cent of Australian households having access to the Internet (up to 97% for those households that have children aged under 15 years), and approximately 50 per cent of Australians are using it to meet their health or healthcare information needs. However, the information seekers find it difficult to express their health information needs as search queries that find the right information, and also the quality, reliability, 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 (if not more important than) high precision [24].

CLEF eHealth[2], established as a lab workshop in 2012 as part of the *Conference and Labs of the Evaluation Forum* (CLEF), has offered evaluation labs

[1] Statistics extracted from the ABS pages at https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8146.0Main+Features12016-17?OpenDocument, titled "8146.0 – Household Use of Information Technology, Australia, 2016–17", on 28 May 2019.

[2] http://clef-ehealth.org/ (last accessed on 28 May 2019).

since 2013 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 data sets to work with and validate the outcomes. More specifically, these labs and their subsequent workshops target (1) developing processing methods and resources 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; and (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.

The CLEF eHealth labs are open for everybody. We particularly welcome academic and industrial researchers, scientists, engineers, and graduate students in natural language processing, machine learning, and biomedical/health informatics to participate. We also encourage participation by multi-disciplinary teams that combine technological skills with biomedical expertise.

This, the seventh year of the evaluation lab (and eight year of the workshop), aiming to build upon the resource development and evaluation approaches by the previous six or seven years of CLEF eHealth [8,9,14,16,26,28,29], offered the following two tasks [15]:

- *Task 1.* Multilingual Information Extraction: *International Classification of Diseases, Version 10* (ICD-10) coding of *non-technical summaries* (NTSs) of animal experiments in German [22] and
- *Task 2. Technology Assisted Reviews* (TAR) in Empirical Medicine in English [13].

In addition, Task 3. Consumer Health Search in Mono- and Multilingual Settings was initially advertised, but unfortunately, due to unforeseen circumstances, it had to be postponed[3].

The *Multilingual Information Extraction* task challenged participants to index German NTSs of animal experiments with the ICD-10 terminology of diseases. A detailed analysis based on the diseases addressed by the NTSs allows more transparency of the animal experiments being carried out by researchers [2]. It could be treated as a text classification or cascaded named entity recognition and normalization task. Even though we only addressed one language (German), we encouraged participants to explore multilingual approaches. The results of high performing systems could be used within the workflow of institutes mandated by the *European Union* (EU) to publish the NTSs approved in their states. The 2019 Task 1 built upon the 2016–2018 information extraction tasks [19–21], which already addressed the ICD-10 terminology to code causes of death from a corpus of death reports in French (2016, 2017, and 2018), English (2017), Hungarian (2018), and Italian (2018). Prior to this, the CLEF eHealth tasks considered *Unified Medical Language System* (UMLS) and *Systematized*

[3] The organizers apologize to the teams that registered their interest in the task for any inconvenience caused by this delay.

Nomenclature of Medicine—Clinical Terms (SNoMed-CT) codification of clinical reports in English in 2013, and UMLS named entity recognition of clinical reports in French in 2015, among others [27].

The *TAR* task 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 results of the explored approaches in the submitted systems towards generating a clear overview of the current scientific consensus could be informing health care and its policy making in the future. This automated generator might release scientists and policy advisors' time from the currently laborious iterative process of conducting publication searches and revising them in order to retrieve all the documents that are relevant for the purposes of writing reliable systematic reviews; this hard challenge is known in the IR domain as the total recall problem and with the number of published medical papers expanding rapidly, the need for automation in this process becomes of utmost importance.

This year's Task 2, differed from the past two years [11,12] by diversifying the focus across different type of reviews including *Diagnostic Test Accuracy* (DTA), *Intervention*, *Prognosis*, and *Qualitative* reviews. Even though search in the area of DTA reviews is generally considered the hardest [18], this year we wanted to investigate how the technology that has been developed over the past two years would extend to other types of reviews. The typical process of searching for scientific publications to conduct a systematic review consists of three stages: (a) specifying a number of inclusion criteria that characterize the articles relevant to the review and constructing a complex Boolean Query to express them, (b) screening the abstracts and titles that result from the Boolean query, and (c) reading and screening the full documents that passed the Abstract and Title Screening. Building on the 2017 task, which focused on the second stage of the process, that is, Abstract and Title Screening, and same as the 2018 task, the 2019 task focused both on the first stage (*subtask 1*) and second stage (*subtask 2*) of the process, that is, Boolean Search and Abstract and Title Screening.

More precisely, these subtasks of Task 2 were defined as follows:

- *Subtask 1.* Prior to constructing a Boolean Query researchers have to design and write a search protocol that in written and in detail defines what constitutes a relevant study for their review. For the challenge associated with the first stage of the process, participants were provided with the relevant pieces of a protocol, in an attempt to complete search effectively and efficiently bypassing the construction of the Boolean query.
- *Subtask 2.* Given the results of the Boolean Search from stage 1 as the starting point, participants were required to rank the set of *abstracts* (A). The task had the following two goals: (i) to produce an efficient ordering of the documents, such that all of the relevant abstracts are retrieved as early as possible, and (ii) to identify a subset of A which contains all or as many of the relevant abstracts for the least effort (i.e., total number of abstracts to be assessed).

The *Consumer Health Search* task was advertised as a continuation of the previous CLEF eHealth IR tasks that ran every year since the onset of

CLEF eHealth evaluation labs in 2013 [5–7,10,23,25,30], and embraced the *Text REtrieval Conference* (TREC) -style evaluation process, with a shared collection of searchable documents and their search queries, the contribution of runs from participants, and the subsequent formation of relevance assessments and evaluation of these participants' submissions. For the first time, the search queries (and their variants) were intended to not only be in written format but also in spoken format, with automatic speech-to-text transcripts provided. The new document collection introduced in the 2018 Task 3, consisting of over 5 million pages from the *World Wide Web* (WWW) was to be used for this task. This was a compilation of Web pages of selected domains acquired from the CommonCrawl[4]. User stories for search query and query variant generation were those, using the discharge summaries and forum posts, we used in previous years of the task.

The remainder of this overview paper is structured as follows: First, in Sect. 2, we detail for each task its text documents; human annotations, queries, and relevance assessments; and evaluation methods. After this, in Sect. 3, we describe the task submissions and results of the CLEF eHealth 2019 evaluation lab. Finally, in Sect. 4 we conclude the study.

2 Materials and Methods

In this section, we describe the materials and methods used in the two tasks of the CLEF eHealth evaluation lab 2019. After specifying our text documents to process in Sect. 2.1, we address their human annotations, queries, and relevance assessments in Sect. 2.2. Finally, in Sect. 2.3 we introduce our evaluation methods. We also include in Sects. 2.1 and 2.2 a brief description of the document set and its intended query set for Task 3.

2.1 Text Documents

Task 1. The multilingual information extraction task challenged its participants with the fully automated semantic indexing of NTSs of animal experiments using codes from the German version of the ICD-10. The NTPs were short publicly-available summaries[5] written as part of the approval procedure for animal experiments in Germany. The database currently contains more than 10, 000 NTPs (as of May/2019).

Task 2. The technologically assisted reviews in empirical medicine task used the PubMed document collection for its Boolean Search challenge and a subset of PubMed documents for its challenge to make Abstract and Title Screening more effective. More specifically, for the Abstract and Title Screening subtask the *PubMed Document Identifiers* (PMIDs) of potentially relevant

[4] http://commoncrawl.org/ (last accessed on 28 May 2019).

[5] The *AnimalTestInfo* database was publicly available at https://www.animaltestinfo.de when the task was launched.

PubMed Document abstracts were provided for each training and test topic. The PMIDs were collected by the task coordinators by re-running the MEDLINE Boolean query used in the original systematic reviews conducted by Cochrane to search PubMed.

Task 3. The document corpus is the same as the corpus used in 2018. It consists of web pages acquired from the CommonCrawl. An initial list of websites was identified for acquisition. The list was built by submitting the CLEF 2018 queries to the Microsoft Bing Apis (through the Azure Cognitive Services) repeatedly over a period of a few weeks, and acquiring the URLs of the retrieved results. The domains of the URLs were then included in the list, except some domains that were excluded for decency reasons. The list was further augmented by including a number of known reliable health websites and other known unreliable health websites, from lists previously compiled by health institutions and agencies.

2.2 Human Annotations, Queries, and Relevance Assessments

Task 1. The task consisted of assigning codes with respect to chapters or groups of the 2016 German Modification of ICD-10[6]. The training and development data set[7] contained a total of 8, 386 NTSs of animal experiments recently carried out in Germany (as of September 2018). It was split into training and development sets with 7, 544 and 842 NTSs, respectively. For the test set, we released 407 NTSs[8] for which participants should predict the ICD-10 codes. In all data sets, each NTS contained a title, benefits (goals) of the experiments, possible harms caused to the animals, and comments related to the *replacement, reduction and refinement* (3R) principles. All documents were in the German language. The data set included the ICD-10 codes manually assigned by experts. However, some NTSs had no ICD-10 codes assigned to them, since the codes were not applicable to the benefits described in the NTS.

Task 2. In Task 2 Subtask 1, for the No-Boolean-Search challenge as input for each topic participants were provided with

1. a Topic-ID,
2. the title of the review, written by Cochrane experts,

[6] Available at https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2016/.

[7] Publicly available on 24 January 2019 at https://www.openagrar.de/receive/openagrar_mods_00046540?lang=en under the *Creative Commons, Attribution-NonCommercial-NoDerivatives 4.0 International* (CC BY-NC-ND 4.0) license as DOI https://doi.org/10.17590/20190118-134645-0.

[8] Publicly available on 6 May 2019 https://www.openagrar.de/receive/openagrar_mods_00049062?lang=en under the *Creative Commons, Attribution-NonCommercial-NoDerivatives 4.0 International* (CC BY-NC-ND 4.0) license.

3. the most important parts of the protocol, written by Cochrane experts, and
4. the entire PubMED database (which was available for downloaded directly from PubMED, through ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline).

In Task 2 Subtask 2, focusing on title and abstract screening, topics consisted of the Boolean Search from the first step of the systematic review process. Specifically, for each topic the following information was provided.

1. a Topic-ID,
2. the title of the review, written by Cochrane experts,
3. the Boolean query, manually constructed by Cochrane experts, and
4. the set of PMIDs returned by running the query in MEDLINE.

Participants were provided with eight topics of DTA reviews, 20 topics of Intervention reviews, one topic of Prognosis, and two of Qualitative reviews, as a test set for both subtasks. The 72 DTA topics (which excludes topics that were reviewed and found unreliable) considered in CLEF 2017 and 2018 TAR tasks were used as training set. Further, we developed 20 Intervention topics that were also provided as training set to participants.

The original systematic reviews written by Cochrane experts included a reference section that listed Included, Excluded, and Additional references to medical studies. The union of Included and Excluded references are the studies that were screened at a Title and Abstract level and were considered for further examination at a full content level. These constituted the relevant documents at the abstract level, while the Included references constituted the relevant documents at the full content level. References in the original systematic reviews were collected from a variety of resources, not only MEDLINE. Therefore, studies that were cited but did not appear in the results of the Boolean query were excluded from the label set for both Subtask 1 and Subtask 2.

Regarding Subtask 2, that is, the Title and Abstract Screening, relevance was assessed at two levels, at abstract level, which expresses the potential of the article to be relevant and included in the review, and hence need to be read in full, and at full content level, after the full article has been read and decided whether to be included or excluded from the study. The following numbers present for each type of study the percentage of relevant document (abstract or content level) in the development set and in the test set, so that the reader can get an idea of the difficulty of the task, the differences across different types of reviews if any, and any changes in the relevance distribution between training and test sets.

Hence, the percentage of relevant document (1) for the DTA studies, (1a) at abstract level, in the training set was 1.7% and in the test set 1.4% of the total number of PMIDs released, while (1b) at content level it was 0.3% in the training set, and 0.8% in the test set. (2) For the Intervention studies, the percentage of relevant documents (2a) at abstract level in the training set was 1.7% and in the test set 0.9%, while at the content level the average percentage was 2.2% in the training set, and 1.2% in the test set. For the Prognosis and Qualitative reviews no training data was provided. (3) In the test set for the Prognosis, (3a) the percentage of relevant documents is 5.7% at the abstract level and (3b) 2.7% at the content level, while (4) for the Qualitative, (4a) the percentage of relevant documents is 1.7% at the abstract level and(4b) 0.4% at the content level.

All the released data for the 2017 – 2019 CLEF eHealth TAR tasks can be found at https://github.com/CLEF-TAR.

Task 3. With the aim to acquire more relevance assessments and increase the collection reusability, the intent this year was to reuse the same set of 50 query narratives developed in 2018's Task 3 [10]. In 2018, query creators devised 7 query variants from each query narrative. This was accomplished by asking laypeople and medical experts to generate written queries based on the textual narratives. In 2019, in order to increase the variability of generated queries, written narratives were converted into spoken audio. After hearing the narratives, a set of query creators were to generate spoken query variants by speaking their queries aloud. Our intention was to make the generated original spoken queries as well as the output of a speech-recognition software available to the participants.

2.3 Evaluation Methods

Task 1. The training and development sets were released on 24 January 2019, and the test set on 6 May 2019. Teams could submit by 13 May 2019 up to three runs/solutions for the test data set. We evaluated the runs based on the usual metrics of the precision, recall, and F-measure using a publicly-available Python script[9].

Task 2. Teams could submit an unlimited number of runs per task. In addition, participants were also encouraged to submit any number of runs that result from their 2017 and 2018 frozen systems. System performance was assessed using the same evaluation approach as that used for the 2018 TAR challenge [12]. Specifically, (i) similarly to the previous year, runs were evaluated on the basis of identifying the studies to be included (relevant documents), (ii) different from previous years, runs were evaluated on the basis of not only finding the studies to be included, but also finding high quality included studies before low quality included studies.

The assumption behind this evaluation approach (i) was the following: The user of your system is the researcher that performs the abstract and title screening of the retrieved articles. Every time an abstract is returned (i.e., ranked) there is an incurred cost/effort, while the abstract is either irrelevant (in which case no further action will be taken) or relevant (and hence passed to the next stage of document screening) to the topic under review.

[9] https://github.com/mariananeves/clef19ehealth-task1.

Evaluation measures were as follows: Area under the recall-precision curve (i.e., Average Precision); Minimum number of documents returned to retrieve all relevant documents; Work Saved over Sampling at different Recall levels; Area under the cumulative recall curve normalized by the optimal area; Recall @ 0% to 100% of documents shown; a number of newly constructed cost-based measures; and reliability [3].

Evaluation approach (ii) considered not only the relevance but the quality of the articles as well, taking into account indicators such as the risk-of-bias, and the sample size of the trials reported of the studies. This second evaluation approach depended on assessments Cochrane reviewers made manually on aspects of the included studies. Obtaining these assessments turned out to be a difficult task therefore this second evaluation approached was postponed for the future.

The training data set was released at the end of March 2019 and the test data set on 14 May 2019. The relevance labels on the testing data (required by active learning techniques) were provided to participants on 14 May 2019 as well, while the submission deadline was set to 21 May 2019 so that participants could not tune their systems towards the actual labels.

More details on the evaluation are provided in the Task 2 overview paper [13].

3 Results

The number of people who registered their interest in CLEF eHealth tasks was 31 in Task 1 and 36 in Task 2. In total, nine teams submitted to the two shared tasks.

Task 1 received considerable interest with the submission of 14 runs from six teams. We had two teams from Germany (MLT-DFKI and WBI), one from India (SSN_NLP), one from Italy (IMS_UNIPD), one as a collaboration between Spain and UK (TALP_UPC) and one from Turkey (DEMIR). Table 1 summarizes the results obtained by each team.

Participants relied on a diverse range of approaches. WBI utilized the multilingual version of the BERT-Base model [4] and made use of additional resources, such as the *German Clinical Trials Registe*r (DRKS)[10]. MLT-DFKI utilized Google Translate to convert documents into English and then relied on pre-trained BioBERT [17] to perform the prediction of ICD-10 codes. DEMIR utilized ElasticSearch for searching for similar NTSs and selected top documents (NTSs) based on *k-nearest neighbors* (KNN) and on threshold-based methods. SSN_NLP relied on a seq2seq mapping model based on bidirectional *long short-term memory* (LSTM) and experimented with the Normed_Bahdanau and the Scaled Luong attention mechanisms. IMS-UNIPD tried three Naïve Bayes classifiers (Bernoulli, Multinomial and Poisson) based on a 2D representation of the probabilities.

[10] See https://www.drks.de/drks_web/setLocale_EN.do.

Task 2 attracted the interest of 3 teams submitting runs, all from Europe, including one team from The Netherlands (UvA), one team from the UK (Sheffield), and one team from Italy (UNIPD). For Subtask 1, we received no runs. For Subtask 2, we received 36 runs from the three teams. The results on a selected subset of metrics on DTA, Intervention, Prognosis, and Qualitative studies are shown in Tables 2, 3, 4, and 5, respectively. The three teams used a variety of ranking methods including traditional BM25, interactive BM25, continuous active learning, relevance feedback, as well as a variety of stopping criteria to provide a threshold on the ranking.

Table 1. System performance for ICD-10 coding on the test set for German NTSs in terms of Precision (P), recall (R) and F-measure (F). The results are ordered in decreasing order of the scores for F-Measure. We highlight in **bold** the highest scores for P, R, and F.

Team	P	R	FM
WBI-run1	0.83	0.77	**0.80**
WBI-run2	**0.84**	0.74	0.79
WBI-run3	0.80	0.78	0.79
MLT-DFKI	0.64	**0.86**	0.73
DEMIR-run1	0.46	0.50	0.48
DEMIR-run3	0.46	0.49	0.48
DEMIR-run2	0.49	0.44	0.46
TALP_UPC	0.37	0.35	0.36
SSN_NLP-run2	0.19	0.27	0.23
SSN_NLP-run1	0.19	0.27	0.22
SSN_NLP-run3	0.13	0.34	0.19
IMS_UNIPD-run3	0.10	0.05	0.07
IMS_UNIPD-run2	0.009	0.50	0.017
IMS_UNIPD-run1	0	0	0