CLEF 2017 eHealth Evaluation Lab Overview

Lorraine Goeuriot[1], Liadh Kelly[2], Hanna Suominen[3], Aurélie Névéol[4], Aude Robert[5], Evangelos Kanoulas[6], Rene Spijker[7], João Palotti[8], and Guido Zuccon[9] *

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France, Lorraine.Goeuriot@imag.fr
[2] ADAPT Centre, Dublin City University, Ireland liadh.kelly@dcu.ie
[3] The Australian National University (ANU), Data61/CSIRO, University of Canberra, and University of Turku, Canberra, ACT, Australia, hanna.suominen@anu.edu.au
[4] LIMSI CNRS UPR 3251 Université Paris-Saclay F-91405 Orsay - France, Aurelie.Neveol@limsi.fr
[5] INSERM - CépiDc 80 rue du Général Leclerc 94276 Le Kremlin-Bicêtre Cedex, France, aude.robert@inserm.fr
[6] Informatics Institute, University of Amsterdam, Netherlands, E.Kanoulas@uva.nl
[7] Cochrane Netherlands and UMC Utrecht, Julius Center for Health Sciences and Primary Care, Netherlands, R.Spijker-2@umcutrecht.nl
[8] Vienna University of Technology, Austria, palotti@ifs.tuwien.ac.at
[9] Queensland University of Technology, Brisbane, QLD, Australia, g.zuccon@qut.edu.au

Abstract. In this paper we provide an overview of the fifth edition of the CLEF eHealth evaluation lab. CLEF eHealth 2017 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 eHealth information in a multilingual setting. This year's lab offered three tasks: Task 1 on multilingual information extraction to extend from last year's task on French corpora, Task 2 on technologically assisted reviews in empirical medicine as a new pilot task, and Task 3 on patient-centered information retrieval (IR) building on the 2013-16 IR tasks. In total 32 teams took part in these tasks (11 in Task 1, 14 in Task 2, and 7 in Task 3). We also continued the replication track from 2016. Herein, we describe the resources created for these tasks, evaluation methodology adopted and 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, Information Extraction, Medical Informatics, Systematic Reviews, Test-set Generation, Text Classification, Text Segmentation, Self-Diagnosis

* In alphabetical order by surname, LG, LK & HS co-chaired the lab. AN & AR , EK & RS , and JP & GZ led Tasks 1–3, respectively.

1 Introduction

This paper presents an overview of the CLEF eHealth 2017 evaluation lab, organized within the Conference and Labs of the Evaluation Forum (CLEF) to support the development of approaches for helping patients, their next-of-kins, and clinical staff in understanding, accessing, and authoring health information in a multilingual setting. This fifth year of the evaluation lab aimed to build upon the resource development and evaluation approaches offered in the previous four years of the lab [19,11,5,10], which focused on patients and their next-of-kins' ease in understanding and accessing health information.

Task 1 addressed *Multi-lingual Information Extraction* (IE) related to diagnosis coding in written text with a focus on unexplored languages corpora, specifically French. English was also offered. This built upon the 2016 task, which analyzed French biomedical text with the IE of causes of death from a corpus of French death reports [15]. This is an essential task in epidemiology, as the determination and analysis of causes of death at a global level informs public health policies. This task was treated as a named entity recognition and normalization task or as a text classification task. Each language could be considered independently, but we encouraged participants to explore multilingual approaches and approaches which could be easily adapted to a new language. Only fully automated means were allowed, that is, human-in-the-loop approaches were not permitted.

Task 2 on *Technology Assisted Reviews in Empirical Medicine* was introduced for the first time in 2017. It was a high-recall Information Retrieval (IR) task that aimed at evaluating search algorithms that seek to identify all studies relevant for conducting a systematic review in empirical medicine. Evidence-based medicine has become an important strategy in health care and policy making. In order to practice evidence-based medicine, it is important to have a clear overview of the current scientific consensus. These overviews are provided in systematic review articles, that summarize all evidence that is published regarding a certain topic (e.g., a treatment or diagnostic test). In order to write a systematic review, researchers have to conduct a search that will retrieve all the documents that are relevant. This is a difficult task, known in the IR domain as the total recall problem. With the reported medical studies expanding rapidly, the need for automation in this process becomes of utmost importance. CLEF 2017 Task 2 had a focus on Diagnostic Test Accuracy (DTA) reviews. Search in this area is generally considered the difficult, and a breakthrough in this field would likely be applicable to other areas as well [12]. The task coordinators considered all 57 systematic reviews conducted by Cochrane[10] experts on DTA studies and published in the Cochrane library[11]. The coordinators of the task managed to reconstruct the MEDLINE Boolean query used for 50 of these systematic reviews. The corpus considered was Document Abstracts and Titles retrieved by these 50 Boolean queries from the medline database (either

[10] `http://www.cochrane.org/`
[11] `http://www.cochranelibrary.com`

through Ovid[12] or PubMed[13]). The goal of the participants was to (a) rank the documents returned by the Boolean query studies, and (b) find an optimal threshold that could inform experts when to stop examining documents in the ranked list. Recall, precision, effort, cost of missing studies, and combinations of these metrics were used to assess the quality of the participating systems. The set of relevant titles and abstracts used in the evaluation were directly extracted from the reference section of the systematic reviews.

Task 3, the IR Task, aimed at evaluating the effectiveness of IR systems when searching for health content on the web, with the objective to foster research and development of search engines tailored to health information seeking. This year's IR task continued the growth path identified in 2013, 2014, 2015, and 2016's CLEF eHealth IR challenges [3,4,16,22]. The corpus (ClueWeb12) and the topics used are similar to 2016's. This year new use cases were explored and the pool of assessed documents deepened. The subtasks within the IR challenge were similar to 2016's: ad hoc search, query variation, and multilingual search. A new subtask was also organized, aimed at exploring methods to personalize health search. Query variations were generated based on the fact that there are multiple ways to express a single information need. Translations of the English queries into several langues were also provided. Participants were required to translate the queries back to English and use the English translation to search the collection.

This paper is structured as follows: in Section 2 we detail the tasks, evaluation and datasets created; in Section 3 we describe the submission and results for each task; and in Section 4 we provide conclusions.

2 Materials and Methods

2.1 Text Documents

Task 1 used a corpus of death certificates comprising free-text descriptions of causes of death as reported by physicians in the standardized causes of death forms in France and in the United States. Each document was manually coded by experts with ICD-10 per international WHO standards. The languages of the challenge this year are French and English. Table 1 below provides some statistics on the datasets.

The new technologically assisted reviews in empirical medicine task, Task 2, used a subset of PubMed documents for its challenge to make Abstract and Title Screening more effective. More specifically the PubMed Document Identifiers (PIDs) of potentially relevant PubMed Document abstracts were provided for each training and test topic. The PIDs 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. A distribution of the number of documents to be ranked by participants per topic can be found in Fig. 1.

[12] `http://ovid.com/site/catalog/databases/901.jsp`
[13] `https://www.ncbi.nlm.nih.gov/pubmed/`

Table 1. Descriptive statistics of the Causes of Death Certificates Corpus

	FR			EN	
	Train (2006–2012)	Dev (2013)	Test (2014)	Train (2015)	Test (2015)
Documents	65,844	27,850	31,690	13,330	6,665
Tokens	1,176,994	496,649	599,127	88,530	40,130
Total ICD codes	266,808	110,869	131,426	39,334	18,928
Unique ICD codes	3,233	2,363	2,527	1,256	900
Unique unseen ICD codes	3,233	224	266	1,256	157

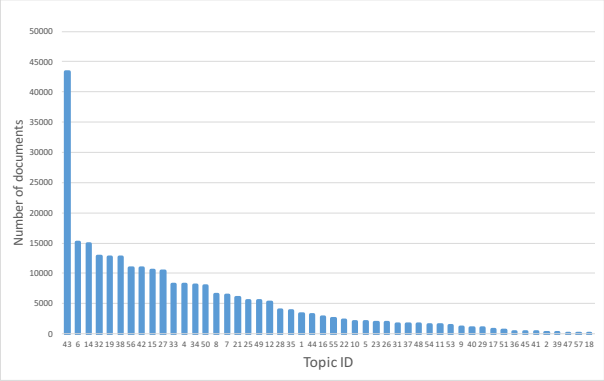

Fig. 1. The distribution of the number of documents across topics in Task 2.

The IR challenge, Task 3, once again used the ClueWeb12 B13[14] corpus, first introduced to the CLEF eHealth IR task in 2016. This corpus is a large snapshot of the Web, crawled between February and May 2012. Unlike the Khresmoi dataset [6] used in earlier years of the IR task [3,4,16], ClueWeb12 does not contain only Health On the Net certified pages and pages from a selected list of known health domains, making the dataset more in line with the material current web search engines index and retrieve. ClueWeb12 B13 contains approximately 52.3 million web pages, for a total of 1.95 TB of data, once uncompressed. For participants who did not have access to the ClueWeb dataset, Carnegie Mellon University granted the organisers permission to make the dataset available through cloud computing instances provided by Microsoft Azure[15]. The Azure instances that were made available to participants for the IR challenge included (1) the Clueweb12 B13 dataset, (2) standard indexes built with the

[14] `http://lemurproject.org/clueweb12/index.php`

[15] The organisers are thankful to Carnegie Mellon University, and in particular to Jamie Callan and Christina Melucci, for their support in obtaining the permission to redistribute ClueWeb 12. The organisers are also thankful to Microsoft Azure who provided the Azure cloud computing infrastructure that was made available to participants through the Microsoft Azure for Research Award CRM:0518649.

Terrier [13] and the Indri [18] toolkits, (3) additional resources such as a spam list [2], Page Rank scores, anchor texts [7], and urls, made available through the ClueWeb12 website.

2.2 Human Annotations, Queries, and Relevance Assessments

For Task 1, the ICD10 codes were abstracted from the raw lines of death certificate text by professional curators at INSERM over the period of 2006-2014 for the French dataset, and curators at the CDC (Center for Disease Control) in the year 2015 for the American dataset. During this time, curators from both groups also manually built dictionaries of terms associated with ICD10 codes. Several versions of these lexical resources were supplied to participants in addition to the training data. Because of the interface used by curators to perform coding, the data used in the challenge comes in separate files: one file contains the original "raw" text of the death certificates presented line by line, one contains the metadata associated with the certificates at the document level, and one contains the ICD codes assigned to the certificate. As detailed in the task overview, a "raw" version of datasets was distributed for French and English. For French, we also distributed an "aligned" version of the data where the ICD10 codes are reconciled with the specific text line that supported the assignment.

For the technology assisted reviews in empirical medicine task, 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. Topic-ID
2. The title of the review, written by Cochrane experts;
3. The Boolean query manually constructed by Cochrane experts;
4. The set of PubMed Document Identifiers (PID's) returned by running the query in MEDLINE.

Twenty of these topics were randomly selected to be used as a training set, while the remaining thirty were used as a test set. 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.

The IR task, Task 3, uses 2016's task topics [22], with the aim to acquire more relevance assessments and improve the collection reusability. The queries consider real health information needs expressed by the general public through posts published in public health web forums. Forum posts were extracted from

the 'askDocs' section of Reddit[16], and presented to query creators. Query creators were asked to formulate queries based on what they read in the initial user post. Six query creators with different medical expertise were used for this task. This year, apart from the AdHoc retrieval task (IRTask 1), the query variation task (IRTask 3) introduced in 2016 and the multilingual task (IRTask 4) introduced in 2013, we proposed a personalized search task (IRTask2) in which participants have to personalize the retrieved list of search results so as to match user expertise, measured by how likely the person is to understand the content of a document (with respect to the health information).

Relevance assessments were collected by pooling participants' submitted runs as well as baseline runs. Assessment was performed by paid medical students who had access to the queries, to the documents, and to the relevance criteria drafted by a junior medical doctor that guided assessors in the judgment of document relevance. The relevance criteria were drafted considering the entirety of the forum posts used to create the queries; a link to the forum posts was also provided to the assessors. Along with relevance assessments, readability/understandability and reliability/trustworthiness judgments were also collected for the assessment pool; these were used to evaluate systems across different dimensions of relevance [21,20].

2.3 Evaluation Methods

Task 1. Teams could submit up to two runs for the tasks for each language. System performance was assessed by the precision, recall and F-measure for ICD code extraction at the document level for English and both at the line and document level for French. Evaluation measures were computed overall (for all ICD codes) and for a subset of the codes, called external causes of death, which are of specific interest to public health specialists. Two baselines were also implemented by the organizers and one participating team.

After submitting their result files for the IE challenges, participating teams had one extra week to submit the system used to produce them, or a remote access to the system, along with instructions on how to install and operate the system. Participating teams were also invited to act as analysts to attempt replicating results with the submitted systems. The replication work is still ongoing at the time of writing this paper.

Task 2. Teams could submit up to eight official runs. System performance was assessed using a Simple Evaluation approach and a Cost-Effective Evaluation approach. The assumption behind the Simple Evaluation approach is 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. Evaluation

[16] `https://www.reddit.com/r/AskDocs/`

measures were: Area under the recall-precision curve, i.e. Average Precision; Minimum number of documents returned to retrieve all R 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 [1]. The assumption behind the Cost-Effective Evaluation approach is the following: The user that performs the screening is not the end-user. The user can interchangeably perform abstract and title screening, or document screening, and decide what PID's to pass to the end-user. Every time an abstract is returned the user can either (a) read the abstract (with an incurred cost CA) and decide whether to pass this PID to the end-user, or (b) read the full document (with an incurred cost of CA+CD) and decide whether to pass this PID to the end-user, or (c) directly pass the PID to the end user (with an incurred cost of 0), or (d) directly discard the PID and not pass it to the end user (with an incurred cost of 0). For every PID passed to the end-user there is also a cost attached to it: CA if the abstract passed on is not relevant, and CA + CD if the abstract passed on is relevant (that is, we assume that the end-user completes a two-round abstract and document screening, as usual, but only for the PIDs the algorithm and feedback user decided to be relevant). More details on the evaluation are provided in the Task 2 overview paper [9].

Task 3. For IRTask 1 (Ad-Hoc Search), participants could treat each query individually (without grouping variants together) and submit up to 7 ranked runs with up to 1,000 documents per query for all 300 queries. For IRTask 2 (Personalized Search), participants could submit up to 7 ranked runs with up to 1,000 documents per information need. For IRTask 3 (Query Variations), participants could submit results for each group of queries of a post, i.e. up to 7 ranked runs with up to 1,000 documents per information need. For IRTask 4 (Multilingual Search), participants could again treat each query individually (like in IRTask 1), submitting up to 7 ranked runs with up to 1,000 documents per query for all 300 queries for each language (Czech (CS), French (FR), Hungarian (HU), German (DE), Polish (PL) and Swedish (SV)).

The organizers also generated baseline runs and a set of benchmark systems using popular IR models implemented in Terrier and Indri. System evaluation was conducted using precision at 10 (p@10) and normalised discounted cumulative gain [8] at 10 (nDCG@10) as the primary and secondary measures, respectively. Precision was computed using the binary relevance assessments; nDCG was computed using the graded relevance assessments. A separate evaluation was conducted using the multidimensional relevance assessments (topical relevance, understandability and trustworthiness) following the methods in [20]. For all runs, Rank biased precision (RBP)[17] was computed along with the multidimensional modifications of RBP, namely uRBP (using binary topicality relevance and understandability assessments), uRBPgr (using graded topicality relevance and understandability assessments), u+tRBP (using binary topicality relevance,

[17] The persistence parameter p in RBP was set to 0.8.

understandability and trustworthiness assessments) and α-uRBP (using a user expertise parameter α, binary topicality relevance and understandability assessments). More details on this multidimensional evaluation are provided in the Task overview paper [17]. Precision and Mean Average Precision were computed using `trec_eval`; while the multidimensional evaluation (comprising RBP) was performed using `ubire`[18].

3 Results

The number of people who registered their interest in CLEF eHealth tasks was 34, 40, and 43 respectively (and a total of 67 unique teams). In total, 32 teams submitted to the three shared tasks.

Task 1 received considerable interest with 34 registered participants. However, only 11 teams submitted runs, including one team from Australia (UNSW), five teams from France (LIMSI, LIRMM, LITL, Mondeca, and SIBM), two teams from Germany (TUC and WBI), one team from Italy (UNIPD), and one team from Russia (KFU). Five teams also submitted systems to the replication track, and two teams also volunteered to participate in the replication track as analysts. The training datasets were released at the end of January 2017 and the test datasets by 25 April 2017. The ICD-10 coding task submission on French and English death certificates were due by 5 May 2017 and the replication track systems by 12 May 2017.

For the English raw dataset, 9 teams submitted 15 runs (Table 2). For the French raw dataset, 6 teams submitted 7 runs for the raw dataset (Table 3) and 9 runs for the aligned dataset(Table 4). In addition to these official runs, unofficial runs were submitted by the task organizers and by some participants after the test submission deadline[19].

The best performance in official runs was achieved with an F-measure of 0.804 for French and of 0.850 for English. Systems relied both on knowledge based methods, machine learning methods, and sometimes a combination of them. The level of performance observed shows that there is potential for integrating automated assistance in the death certificate coding work flow both in French and in English. We hope that continued efforts towards reproducibility will support the shift from research prototypes to operational production systems. See the Task 1 overview paper for further details [14].

Task 2 also received much interest with 40 registered participants. Of these 14 teams submitted runs, including 1 team from Australia (QUT), 1 team from Canada (Waterloo), 1 team from China (ECNU), 1 team from France (CNRS), 1 team from Greece (AUTH), 1 team from India (IIIT), 1 team from Italy (Padua), 1 team from the Netherlands (AMC), 1 team from Singapore (NTU), 1 team from Switzerland (ETH), 3 teams from the United Kingdom (Sheffield, UCL, UOS), and 1 team from the United States (NCSU). The training datasets were released on the 10 March 2017 and the test datasets (with gold standard

[18] `https://github.com/ielab/ubire`, [20].
[19] See Task 1 paper for details on unofficial runs [14].

Table 2. System performance for ICD10 coding on the English raw test corpus in terms of Precision (P), recall (R), and F-measure (F). The top part of the table displays official runs, while the bottom part displays baseline runs.

	ALL				EXTERNAL			
	Team	P	R	F	**Team**	P	R	F
Official runs submitted	KFU-run1	.893	.811	**.850**	KFU-run1	.584	.357	**.443**
	KFU-run2	.891	**.812**	.850	KFU-run2	.631	.325	.429
	TUC-MI-run1	**.940**	.725	.819	SIBM-run1	.426	.389	.407
	SIBM-run1	.839	.783	.810	LIRMM-run2	.233	**.524**	.323
	TUC-MI-run2	.929	.717	.809	LIRMM-run1	.232	**.524**	.322
	WBI-run1	.616	.606	.611	TUC-MI-run1	.880	.175	.291
	WBI-run2	.616	.606	.611	TUC-MI-run2	**1.00**	.159	.274
	LIRMM-run1	.691	.514	.589	UNSW-run1	.168	.262	.205
	LIRMM-run2	.646	.527	.580	Unipd-run2	.292	.111	.161
	Unipd-run1	.496	.442	.468	WBI-run1	.246	.119	.160
	UNSW-run1	.401	.352	.375	WBI-run2	.246	.119	.160
	Unipd-run2	.382	.341	.360	Unipd-run1	.279	.095	.142
	UNSW-run2	.371	.328	.348	UNSW-run2	.043	.310	.076
	Mondeca-run1	*invalid format*			Mondeca-run1	*invalid format*		
	average	.670	.582	.622	**average**	.405	.267	.261
	median	.646	.606	.611	**median**	.279	.262	.274
	Frequency baseline	.115	.085	.097	Frequency baseline	0.00	0.00	0.00
	ICD baseline	.029	.007	.011	ICD baseline	0.00	0.00	0.00

Table 3. System performance for ICD10 coding on the French raw test corpus in terms of Precision (P), recall (R), and F-measure (F). A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays baseline runs.

	ALL				EXTERNAL			
	Team	P	R	F	**Team**	P	R	F
Official runs	SIBM-run1	**.857**	**.689**	**.764**	SIBM-run1	**.567**	**.431**	**.490**
	LITL-run2	.666	.414	.510	LIRMM-run1	.443	.367	.401
	LIRMM-run1	.541	.480	.509	LIRMM-run2	.443	.367	.401
	LIRMM-run2	.540	.480	.508	LITL-run2	.560	.283	.376
	LITL-run1	.651	.404	.499	LITL-run1	.538	.277	.365
	TUC-MI-run2	.044	.026	.033	TUC-MI-run2	.010	.004	.005
	TUC-MI-run1	.025	.015	.019	TUC-MI-run1	.006	.005	.005
	average	.475	.358	.406	**average**	.367	.247	.292
	median	.541	.414	.508	**median**	.443	.283	.376
	Frequency baseline	.339	.237	.279	Frequency baseline	.381	.110	.170

Table 4. System performance for ICD10 coding on the French aligned test corpus in terms of Precision (P), recall (R), and F-measure (F). A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays baseline runs.

	ALL				EXTERNAL			
	Team	P	R	F	**Team**	P	R	F
Official runs	SIBM-run1	.835	**.775**	**.804**	SIBM-run1	.534	**.472**	**.501**
	WBI-run1	.780	.751	.765	TUC-MI-run2	**.740**	.318	.445
	TUC-MI-run2	**.874**	.611	.719	LIRMM-run1	.412	.403	.407
	LITL-run1	.612	.550	.579	LIRMM-run2	.412	.403	.407
	LIRMM-run1	.506	.530	.518	LITL-run1	.482	.348	.404
	LIRMM-run2	.505	.530	.517	LITL-run2	.534	.275	.363
	LITL-run2	.646	.402	.495	WBI-run1	.709	.151	.249
	TUC-MI-run1	.426	.297	.350	TUC-MI-run1	.218	.119	.154
	average	.648	.555	.593	**average**	.505	.311	.366
	median	.629	.540	.548	**median**	.508	.333	.406
	Frequency baseline	.640	.470	.542	Frequency baseline	.508	.338	.406
	ICD baseline	.346	.041	.073	ICD baseline	.000	.000	.000

annotations) by May 2017. Participants submissions were due by 14 May 2017. In total, 14 teams submitted at least one run. See the Task 2 overview paper for further details and the results of the evaluation [9].

Task 3 received much interest with 43 registered participants. Of these 7 teams submitted runs, including 1 team from Australia (QUT), 1 team from Austria (TUW), 1 team from Botswana (UB-Botswana), 1 team from Czech Republic (CUNI), 1 team from Korea (KISTI), 1 team from Portugal (UEvora), and 1 team from Spain (SINAI). Participants submissions were due by 9 June 2017 and the relevance assessments are being collected at the time of writing of this paper. See the Task 3 overview paper for further details and the results of the evaluation [17].

4 Conclusions

In this paper we provided an overview of the CLEF eHealth 2017 evaluation lab. In recent year's the CLEF eHealth lab has offered a recurring contribution to the creation and dissemination of test collections in the fields of biomedical IR and IE. This edition of CLEF eHealth offered three tasks: Task 1 on multilingual IE to extend from last year's task on French corpora, Task 2 on technologically assisted reviews in empirical medicine as a new pilot task, and Task 3 on patient-centred IR extending the 2013–16 IR tasks. We also continued the replication track from 2016 in Task 1. More specifically, Task 1 offered test collections addressing the task of automatic coding using the International Classification of Diseases for death certificates in two languages. Task 2 and Task 3 offered test collections addressing two aspects of biomedical IR: high-recall

IR over PubMed Abstracts and Titles for the purpose of conducting systematic reviews of Diagnostics Test Accuracy studies (Task 2) and effectiveness, quality, and personalization for health related searches made on the Web (Task 3).

Each task's test collections offered 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. The established CLEF eHealth IE and IR tasks (Task 1 and Task 3) used a traditional shared task model evaluation approach again this year whereby a community-wide evaluation is executed in a controlled setting: participants have access to test data at the same time, following which no further updates to systems are allowed and following submission of the outputs from their frozen IE or IR system to the task organiser, their results are evaluated blindly by an independent third party who reports label results for all participants. With our new pilot IR task (Task 2) we aspire to offering means to conduct cross comparable relevance feedback loops, with plans to introduce a newer form of shared evaluation next year through the use of a live evaluation service.

The CLEF eHealth lab has matured and established its presence during its five iterations in 2013–2017. In total, 67 unique teams registered their interests and 32 teams took part in the 2017 tasks (11 in Task 1, 14 in Task 2, and 7 in Task 3). In comparison, in 2016, 2015, 2014, and 2013, the number of team registrations was 116, 100, 220, and 175, respectively and the number of participating teams was 20, 20, 24, and 53 [19,11,5,10]. Given the significance of the tasks, all test collections and resources associated with the lab have been made available to the wider research community through our CLEF eHealth website[20].

Acknowledgements

The CLEF eHealth 2017 evaluation lab has been supported in part by (in alphabetical order) the ANR, the French National Research Agency, under grant CABeRneT ANR-13-JS02-0009-01, the CLEF Initiative and Data61. We are also thankful to the people involved in the annotation, query creation, and relevance assessment exercise. Last but not least, we gratefully acknowledge the participating teams' hard work. We thank them for their submissions and interest in the lab.

References

1. Cormack, G.V., Grossman, M.R.: Engineering quality and reliability in technology-assisted review. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 75–84. SIGIR '16, ACM, New York, NY, USA (2016), http://doi.acm.org/10.1145/2911451.2911510

[20] https://sites.google.com/site/clefehealth/

2. Cormack, G.V., Smucker, M.D., Clarke, C.L.: Efficient and effective spam filtering and re-ranking for large web datasets. Information retrieval 14(5), 441–465 (2011)
3. 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)
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. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: Medical Informatics Europe 2012 (MIE 2012), Village of the Future (2012)
7. Hiemstra, D., Hauff, C.: Mirex: Mapreduce information retrieval experiments. arXiv preprint arXiv:1004.4489 (2010)
8. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422–446 (2002)
9. Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: Overview of the CLEF technologically assisted reviews in empirical medicine. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2017)
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. Leeflang, M.M., Deeks, J.J., Takwoingi, Y., Macaskill, P.: Cochrane diagnostic test accuracy reviews. Systematic reviews 2(1), 82 (2013)
13. Macdonald, C., McCreadie, R., Santos, R.L., Ounis, I.: From puppy to maturity: Experiences in developing terrier. Proc. of OSIR at SIGIR pp. 60–63 (2012)
14. 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)
15. Névéol, A., Cohen, K.B., 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: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (2016)
16. 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)

17. 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)
18. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on Intelligent Analysis. vol. 2, pp. 2–6. Citeseer (2005)
19. 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)
20. Zuccon, G.: Understandability biased evaluation for information retrieval. In: Advances in Information Retrieval. pp. 280–292 (2016)
21. Zuccon, G., Koopman, B.: Integrating understandability in the evaluation of consumer health search engines. In: Medical Information Retrieval Workshop at SIGIR 2014. p. 32 (2014)
22. 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)