Overview of the CLEF eHealth Evaluation Lab 2018

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.

DOI
10.1007/978-3-319-98932-7_26

Publication date
2018

Document Version
Final published version

Published in
Experimental IR Meets Multilinguality, Multimodality, and Interaction

Overview of the CLEF eHealth Evaluation Lab 2018

Hanna Suominen[1,2(✉)], Liadh Kelly[3], Lorraine Goeuriot[4], Aurélie Névéol[5], Lionel Ramadier[5], Aude Robert[6], Evangelos Kanoulas[7], Rene Spijker[8], Leif Azzopardi[9], Dan Li[7], Jimmy[10], João Palotti[11,12], and Guido Zuccon[10]

[1] University of Turku, Turku, Finland
[2] The Australian National University (ANU), Data61/Commonwealth Scientific and Industrial Research Organisation (CSIRO), University of Canberra, Canberra, ACT, Australia
hanna.suominen@anu.edu.au
[3] Maynooth University, Maynooth, Ireland
liadh.kelly@mu.ie
[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
Lorraine.Goeuriot@imag.fr
[5] LIMSI CNRS UPR 3251 Université Paris-Saclay, 91405 Orsay, France
Aurelie.Neveol@limsi.fr
[6] INSERM - CépiDc 80 rue du Général Leclerc, 94276 Le Kremlin-Bicêtre Cedex, France
aude.robert@inserm.fr
[7] Informatics Institute, University of Amsterdam, Amsterdam, Netherlands
E.Kanoulas@uva.nl
[8] Cochrane Netherlands and UMC Utrecht, Julius Center for Health Sciences and Primary Care, Utrecht, Netherlands
R.Spijker-2@umcutrecht.nl
[9] Computer and Information Sciences, University of Strathclyde, Glasgow, UK
leif.azzopardi@strath.ac.uk
[10] Queensland University of Technology, Brisbane, QLD, Australia
jimmy@hdr.qut.edu.au, g.zuccon@qut.edu.au
[11] Vienna University of Technology, Vienna, Austria
palotti@ifs.tuwien.ac.at
[12] Qatar Computing Research Institute, Doha, Qatar
jpalotti@hbku.edu.qa

Abstract. In this paper, we provide an overview of the sixth annual edition of the CLEF eHealth evaluation lab. CLEF eHealth 2018 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 and English corpora to French, Hungarian, and Italian; Task 2 on technologically assisted reviews in empirical medicine building on last year's pilot task

G. Zuccon—In alphabetical order by forename, HS, LK & LG co chaired the lab. AN & LR & AR, EK & RS & LA & DL, and J & JP & GZ led Tasks 1–3, respectively.

P. Bellot et al. (Eds.): CLEF 2018, LNCS 11018, pp. 286–301, 2018.
https://doi.org/10.1007/978-3-319-98932-7_26

in English; and Task 3 on Consumer Health Search (CHS) in mono- and multilingual settings that builds on the 2013–17 Information Retrieval tasks. In total 28 teams took part in these tasks (14 in Task 1, 7 in Task 2 and 7 in Task 3). Herein, we describe the resources created for these tasks, outline our 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 · Total recall · Test-set generation
Text classification · Text segmentation · Self-diagnosis

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. Medical content is becoming available electronically in a variety of forms ranging from patient records and medical dossiers, scientific publications and health-related websites to medical-related topics shared across social networks. Laypeople, clinicians and policy-makers need to easily retrieve, and make sense of medical 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 aims to bring together researchers working on related information access topics and provide them with datasets to work with and validate the outcomes. The vision for the Lab is two-fold: (1) to develop tasks that potentially impact patient understanding of medical information and (2) to provide the community with an increasingly sophisticated dataset of clinical narrative, enriched with links to standard knowledge bases, evidence-based care guidelines, systematic reviews, and other further information, to advance the state-of-the-art in multilingual information extraction and IR in health care. Furthermore, we aim to support reproducible research by encouraging participants to reflect on methods and practical steps to take to facilitate the replication of their experiments. In particular, each year we call participants to submit their systems and configuration files, and independent researchers to reproduce the results of the participating teams.

This, the sixth year of the lab, aiming to build upon the resource development and evaluation approaches offered in the previous five years of the lab [7,8,13,14,26], offered the following three tasks:

- *Task 1.* Multilingual Information Extraction: *International Classification of Diseases, Version 10* (ICD10) coding of death certificates [21],
- *Task 2.* Technologically Assisted Reviews in Empirical Medicine [12], and
- *Task 3.* Consumer Health Search [10].

The *Multilingual Information Extraction* task challenged participants to information extraction in written text with its focus on unexplored languages corpora, specifically French, Hungarian, and Italian this year. This built upon the 2016 and 2017 tasks [19,20] which already addressed the analysis of French and English biomedical text with the extraction of causes of death from a corpus of death reports in French (2016 and 2017) and English (2017). This task can be treated as a named entity recognition and normalization task, but also as a text classification task. Each language can be addressed independently, but we encouraged participants to explore multilingual approaches. Only fully automated means were allowed, that is, human-in-the-loop approaches were not permitted. The goal of the task was to automatically assign ICD10 codes to the text content of death certificates. The results of high performing systems could be used within the workflow of institutes mandated by the *World Health Organisation* (WHO) to provide national death statistics.

The *Technologically Assisted Reviews in Empirical Medicine* 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. This year's task, similar to last year [11], had a focus on *Diagnostic Test Accuracy* (DTA) reviews. Search in this area is generally considered the hardest, and a breakthrough in this field would likely be applicable to other areas as well [15]. 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) 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, 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. More precisely, these tasks 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* (CHS) task was a continuation of the previous CLEF eHealth IR tasks that ran in 2013, 2014, 2015, 2016, and 2017 [4–6,22,23,27], and embraced the *Text REtrieval Conference* (TREC) -style evaluation process, with a shared collection of documents and queries, the contribution of runs from participants and the subsequent formation of relevance assessments and evaluation of the participants submissions. The 2018 task used a new web corpus and a new set of queries compared to previous years. The subtasks within the IR challenge were similar to 2017's: ad hoc search, query variation, methods to personalize health search, and multilingual search. A new subtask was also introduced this year which required participants to classify queries with respect to the underlying query intent as detailed in [3]. 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 languages were also provided. Participants were required to translate the queries back to English and use the English translation to search the collection.

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

2 Materials and Methods

In this section, we describe the materials and methods used in the three tasks of the CLEF eHealth evaluation lab 2018. 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.

2.1 Text Documents

Task 1. The multilingual information extraction: ICD10 coding of death certificates task challenged its participants to information extraction in written text with focus on unexplored languages corpora, specifically French, Hungarian, and Italian this year to supplement last year's task on French and English. Its data set, called the *CepiDC Causes of Death Corpus*, comprised free-text descriptions of causes of death as reported by physicians in the standardized causes of death forms. Each document was manually coded by experts with ICD10 per international WHO standards.

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 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 used in the Consumer Health Search task consists of web pages acquired from the CommonCrawl[1]. An initial list of websites was identified for acquisition. The list was built by submitting the task queries to the Microsoft Bing APIs (through Azure Cognitive Services) repeatedly over a period of a few weeks to incorporate possibly evolving results and variations in the Bing APIs services [9]; results were acquired as URLs and pooled. 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. We decided to include also known unreliable websites so that the collection can serve also for the study of methods that account for the reliability and trustworthiness of the search results.

2.2 Human Annotations, Queries, and Relevance Assessments

Task 1. The task consisted of extracting ICD10 codes from the raw lines of death certificate text (the process of identifying a single ICD code per certificate as the primary cause of death was not evaluated). This task relied on the text supplied to extract ICD10 codes from the certificates, line by line. The extraction system was to generate the ICD10 codes relevant to assign to each line. Systems were encouraged to report evidence text supporting the ICD10 code recommendations in the form of an excerpt of the original text that supports the ICD code prediction. For French, two data formats were supported. The so-called raw format supplied the text of each certificate line separately from the gold standard codes that were supplied at the certificate level. The so-called aligned format reconciled the gold standard codes to the specific certificate line that yielded them. For the French subtask, a training set of 125,384 death certificates and an independent test set of 11,932 death certificates was annotated with respect to ICD10 codes and supporting text evidence by professional coders. For the Hungarian subtask, a training set of 84,703 death certificates and an independent test set of 21,176 death certificates was assigned ICD10 codes by professional coders. For the Italian subtask, a training set of 14,502 death certificates and an independent test set of 3,618 death certificates was assigned ICD10 codes by professional coders.

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

[1] http://commoncrawl.org/.

1. Topic-ID.
2. The title of the review, written by Cochrane experts.
3. A part of the protocol: The Objective, the Type of Study, the Participants, the Index Tests, the Target Conditions, and the Reference Standards.
4. The entire PubMED database (which was available for downloaded directly from PubMED).

Participants were provided with 30 topics of Diagnostic Test Accuracy (DTA) reviews.

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. 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 (PMID's) returned by running the query in MEDLINE.

The CLEF 2017 TAR 42 topics (which excludes topics that were reviewed and found unreliable) were used as training set. A new test set consisting of 30 topics of Diagnostic Test Accuracy (DTA) reviews was generated for this year's challenge. The total number of unique PMID's released for the training set was 241,669 (an average of 5,754 per topic) and for the test set 218,496 (an average of 7,283 per topic).

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. The average percentage of relevant documents at Abstract level in the training set is 3.8% of the total number of PMID's released, and in the test set 4.7%, while at the content level the average percentage is 1.5% in the training set, and 1% in the test set.

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 Subtask 2, but included for Subtask 1. Hence, the total number of relevant abstracts in the test set for Subtask 1 increased to 4,656 from 3,964 in Subtask 2, and the total number of relevant studies increased to 759 from 678. An important note here is that the additional studies are also included in the MEDLINE database, they were simply not retrieved by the Boolean query.

Task 3. The CHS task, Task 3, uses a new set of 50 queries issued by the general public to the HON search services, manually labeled with search intent and translated into French, German and Czech [3]. Subtask 1 uses these 50 queries. For subtask 2 and 3, each topic is augmented with 6 query variations issued by 6 research students at QUT with no medical knowledge. Each student was asked to formulate a query for each of the 50 queries' narrative. No post-processing was done to the formulated query variations and duplicates might exist within the 6 variations of a query. Subtask 4 uses parallel queries in the following languages: French, German, and Czech. These queries are manual translations of Subtask 1's 50 queries. Subtask 5 contains the same 50 topics labeled with search intents: (1) Disease/illness/syndrome/pathological condition, (2) Drugs and medicinal substances, (3) Healthcare, (4) Test & procedures, (5) First aid, (6) Healthy lifestyle, (7) Human anatomy, (8) Organ systems.

Relevance assessments are currently in progress. Similar to the 2016 and 2017 pools, we created the pool using the RBP-based Method A (Summing contributions) by Moffat et al. [17], in which documents are weighted according to their overall contribution to the effectiveness evaluation as provided by the RBP formula (with $p = 0.8$, following Park and Zhang [24]). This strategy, named RBPA, was chosen because it was shown that it should be preferred over traditional fixed-depth or stratified pooling when deciding upon the pooling strategy to be used to evaluate systems under fixed assessment budget constraints [16], as it is the case for this task.

Along with relevance assessments, readability/understandability and reliability/trustworthiness judgments will also be collected for the assessment pool; these will be used to evaluate systems across different dimensions of relevance. We plan to use crowdsourcing for the acquisition of the relevance assessments.

2.3 Evaluation Methods

Task 1. After completing our data use agreement, authorized participants were able to obtain training sets from March 2018. The test data for CLEF eHealth 2018 Task 1 was released on 27 April 2018. Teams could submit up to 2 runs per dataset by 12 May 2018. Hence, the maximum was 8 runs for all four datasets. System performance was assessed by the precision, recall and F-measure for ICD code extraction at the document level for Hungarian and Italian and both at the line and document level for French. Evaluation measures were computed overall for all ICD codes. A baseline was also implemented by the organizers [21].

Task 2. Teams could submit up to 3 runs per task. Hence a maximum of 6 runs for both subtasks. In addition, for Subtask 2, participants were also encouraged to submit ANY number of runs that result from their 2017 frozen systems. System performance was assessed using the same evaluation approach as that used for the 2017 TAR challenge [11]. The assumption behind this 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 measures were: Area under the recall-precision curve, that is, 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]. More details on the evaluation are provided in the Task 2 overview paper [12].

Task 3. For Subtasks 1, 2, and 3, participants could submit up to 4 runs in TREC format. For Subtask 4, participants could submit up to 4 runs per language. For Subtask 5, teams could submit runs containing up to 3 candidate intent per query, with up to 4 variation run. Evaluation measures for Subtasks 1 and 4 were NDCG@10, BPref and RBP. Subtask 2 used uRBP (with alpha value capturing the user expertise). Subtask 3 used NDCG@10, BPref and RBP - in the MVE framework. For Subtask 5, the evaluation measures are Mean Reciprocal Rank, nDCG@1, 2, 3.

3 Results

The number of groups who registered their interest in CLEF eHealth tasks was 26, 42, and 46 respectively (and a total of 70 unique teams). In total, 28 teams submitted to the three shared tasks.

Task 1 received considerable interest with 14 teams submitting runs, including one team from Algeria (techno), one team from Canada (TorontoCL), two teams from China (ECNU and WebIntelligentLab), three teams from France (APHP, IAM, ISPED), one team from Germany (WBI), one team from Italy (UNIPD), three teams from Spain (IxaMed, SINAI and UNED), one team from Switzerland (SIB) and one team from the United Kingdom (KCL). The training datasets were released at the beginning of March 2018 and the test datasets by 27 April 2018. The ICD-10 coding task submission on French, Hungarian and Italian death certificates were due by 12 May 2018.

For the Hungarian raw dataset, we received 9 official runs from 5 teams (Table 3). For the Italian raw dataset, we received 12 official runs from 7 teams (Table 4). For the French raw dataset, we received 18 official runs from 12 teams (Table 2). For the French aligned dataset, we received 16 official runs from 8 teams (Table 1). In addition to these official runs, unofficial runs were submitted by some participants after the test submission deadline[2].

Participants relied on a diverse range approaches including classification methods (often leveraging neural networks), information retrieval techniques and

[2] See Task 1 paper for details on unofficial runs [20].

Table 1. 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 the baseline

Team	P	R	F
IxaMed-run2	0.841	**0.835**	**0.838**
IxaMed-run1	**0.846**	0.822	0.834
IAM-run2	0.794	0.779	0.786
IAM-run1	0.782	0.772	0.777
SIB-TM	0.763	0.764	0.764
TorontoCL-run2	0.810	0.720	0.762
TorontoCL-run1	0.815	0.712	0.760
KCL-Health-NLP-run1	0.787	0.553	0.649
KCL-Health-NLP-run2	0.769	0.537	0.632
SINAI-run2	0.733	0.534	0.618
SINAI-run1	0.725	0.528	0.611
WebIntelligentLab	0.673	0.491	0.567
ECNUica-run1	0.771	0.437	0.558
ECNUica-run2	0.771	0.437	0.558
techno	0.489	0.356	0.412
KR-ISPED	0.029	0.020	0.023
Average	0.712	0.581	0.634
Median	0.771	0.545	0.641
Frequency baseline	0.452	0.450	0.451

dictionary matching accommodating for different levels of lexical variation. Most participants (12 teams out of 14) used the dictionaries that were supplied as part of the training data as well as other medical terminologies and ontologies (at least one team).

Task 2 attracted the interest of 7 teams submitting runs, including one team from Canada (UWA), one team from the USA (UIC/OHSU), one team from the UK (Sheffield), one team from China (ECNU), one team from Greece (AUTH), one team from Italy (UNIPD), one team from France (Limsi-CNRS). For the subtask 1, we received 12 runs from 4 teams. The results on a selected subset of metrics are shown in Table 5. For the subtask 2, we received 19 runs from 7 teams. The results on a selected subset of metrics are shown in Table 6. The 7 teams used a variety of learning methods including batch supervised learning, continuous active learning, a variety of learning algorithms including logistic regression, support vector machines, and neural networks, as well as unsupervised retrieval methods, such as TT-IDF, BM25, with or without traditional relevance feedback

Table 2. 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 the baseline.

Team	P	R	F
IxaMed-run1	0.872	**0.597**	**0.709**
IxaMed-run2	0.877	0.588	0.704
LSI-UNED-run1	0.842	0.556	0.670
LSI-UNED-run2	0.879	0.540	0.669
IAM-run2	0.820	0.560	0.666
IAM-run1	0.807	0.555	0.657
TorontoCL-run2	0.842	0.522	0.644
TorontoCL-run1	0.847	0.515	0.641
WebIntelligentLab	0.702	0.495	0.580
ECNUica-run1	0.790	0.456	0.578
KCL-Health-NLP-run1	0.738	0.405	0.523
KCL-Health-NLP-run2	0.724	0.394	0.510
ims-unipd	0.653	0.396	0.493
techno	0.569	0.286	0.380
WBI-run2	0.512	0.253	0.339
WBI-run1	0.494	0.246	0.329
KR-ISPED	0.043	0.021	0.028
ECNUica-run2	**1.000**	0.000	0.000
Average	0.723	0.410	0.507
Median	0.798	0.475	0.579
Frequency baseline	0.341	0.201	0.253

methods, such as the Rocchio's Algorithm, and a variety of text representation methods including simple count-based methods to neural embeddings.

The training datasets were released on February 2018 and the test datasets on March 2018. The relevance labels on the testing data (required by active learning techniques) were provided to participants on 1 May 2018, four days before the submission deadline so that participants could not tune their systems towards the actual labels.

Task 3 had seven teams submitting runs: one team from Australia (QUT), one team from Botswana (UB-Botswana), one team from Czech Republic (CUNI), one team from Italy (IMS Unipd), one team from Portugal (UEvora), one team from Spain (SINAI), and one team from Tunisia (MIRACL). Participants submissions were due by June 8th 2018 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 [10].

Table 3. System performance for ICD10 coding on the **Hungarian raw** test corpus in terms of Precision (P), recall (R) and F-measure (F).

Hungarian (raw)			
Team	P	R	F
IxaMed run2	**0.970**	**0.955**	**0.963**
IxaMed run1	0.968	0.954	0.961
LSI UNED-run2	0.946	0.911	0.928
LSI UNED-run1	0.932	0.922	0.927
TorontoCL-run2	0.922	0.897	0.910
TorontoCL-run1	0.901	0.887	0.894
ims unipd	0.761	0.748	0.755
WBI-run2	0.522	0.388	0.445
WBI-run1	0.518	0.384	0.441
Average	0.243	0.174	0.202
Median	0.646	0.606	0.611
Frequency baseline	0.115	0.085	0.097

Table 4. System performance for ICD10 coding on the **Italian raw** test corpus in terms of Precision (P), recall (R) and F-measure (F).

Italian (raw)			
Team	P	R	F
IxaMed run1	**0.960**	**0.945**	**0.952**
IxaMed run2	0.945	0.922	0.934
LSI UNED-run1	0.917	0.875	0.895
LSI UNED-run2	0.931	0.861	0.895
TorontoCL-run1	0.908	0.824	0.864
TorontoCL-run2	0.900	0.829	0.863
WBI-run2	0.862	0.689	0.766
WBI-run1	0.857	0.685	0.761
KCL-Health-NLP-run1	0.746	0.636	0.687
KCL-Health-NLP-run2	0.725	0.616	0.666
ims unipd	0.535	0.484	0.509
Average	0.844	0.761	0.799
Median	0.900	0.824	0.863
Frequency baseline	0.165	0.172	0.169

Table 5. Average scores for the submitted runs in task 2 - subtask 1.

Run	MAP	R@50	R@100	R@200	R@300	R@400	R@500	R@1000	R@2000	R@k
auth_run1	0.113	0.188	0.341	0.51	0.61	0.66	0.693	0.787	0.802	0.816
auth_run2	0.113	0.188	0.341	0.51	0.61	0.66	0.693	0.787	0.802	0.809
auth_run3	0.113	0.188	0.341	0.51	0.61	0.66	0.693	0.787	0.802	0.787
ECNU_RUN1	0.072	0.17	0.242	0.339	0.393	0.431	0.472	0.561	0.561	0.472
ECNU_RUN2	0.041	0.076	0.145	0.216	0.281	0.34	0.378	0.378	0.378	0.378
ECNU_RUN3	0.072	0.173	0.246	0.341	0.411	0.452	0.485	0.561	0.561	0.485
shef-bm25	0.026	0.045	0.063	0.108	0.149	0.169	0.187	0.261	0.315	0.426
shef-tfidf	0.002	0.005	0.005	0.017	0.029	0.042	0.057	0.086	0.126	0.266
shef-bool	0.008	0.022	0.049	0.069	0.097	0.111	0.124	0.17	0.221	0.299
UWA	0.124	**0.256**	**0.428**	**0.592**	**0.693**	**0.771**	**0.806**	**0.912**	**0.947**	0.951
UWX	**0.154**	0.254	0.386	0.564	0.673	0.743	0.784	0.884	0.95	0.951
UWG	0.080	0.121	0.273	0.462	0.59	0.675	0.729	0.883	0.959	**0.962**

Table 6. Average scores for the submitted runs in task 2 - subtask 2.

Run	MAP	R@10%	R@20%	R@30%	R@K	K	Last_Rel	WSS95	WSS100
auth_run1	0.400	0.655	**0.883**	0.943	1.000	7283	3405	0.749	0.611
auth_run2	0.400	0.655	**0.883**	0.943	0.944	880	3405	0.749	0.611
auth_run3	0.393	0.653	0.874	0.931	0.943	880	4295	0.734	0.563
cnrs_RF_bi	0.314	0.560	0.776	0.862	1.000	7283	5173	0.617	0.460
cnrs_comb	0.337	0.557	0.774	0.862	1.000	7283	4378	0.657	0.510
cnrs_RF_uni	0.313	0.554	0.766	0.833	1.000	7283	5708	0.513	0.349
ECNU_RUN1	0.142	0.259	0.462	0.580	0.520	465	7173	0.027	0.026
ECNU_RUN2	0.081	0.232	0.414	0.539	0.371	466	4725	0.019	0.000
ECNU_RUN3	0.146	0.303	0.511	0.614	0.534	465	7172	0.029	0.025
unipd_t1500	0.316	0.544	0.761	0.843	0.945	2188	4259	0.543	0.396
unipd_t1000	0.317	0.542	0.765	0.857	0.920	1600	4101	0.572	0.410
unipd_t500	0.321	0.556	0.786	0.865	0.856	873	3935	0.616	0.475
shef-fb	**0.607**	0.554	0.774	0.856	1.000	7283	5171	0.635	0.444
shef-general	0.258	0.373	0.635	0.773	1.000	7283	5519	0.552	0.431
shef-query	0.224	0.338	0.591	0.734	1.000	7283	5736	0.506	0.377
uci_model8	0.174	0.289	0.462	0.562	0.513	1752	6385	0.255	0.154
uic_model7	0.180	0.296	0.473	0.579	0.576	2120	6185	0.264	0.164
UWB	0.378	**0.656**	**0.883**	0.944	0.927	1764	2655	0.756	0.610
UWA	0.362	0.651	0.877	**0.945**	0.990	2926	**2545**	0.751	0.608

4 Conclusions

In this paper, we provided an overview of the CLEF eHealth 2018 evaluation lab. The CLEF eHealth workshop series was established in 2012 as a scientific workshop with an aim of establishing an evaluation lab [25]. Since 2013, this annual workshop has been supplemented with two or more preceding shared tasks each year, in other words, the CLEF eHealth 2013–2018 evaluation labs [7,8,13,14,26]. During these past seven years, 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.

Test collections generated by each of the three CLEF eHealth 2018 tasks 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 for evaluation in which a community-wide evaluation is executed in a controlled setting: independent training and test datasets are used and all participants gain access to the test data at the same time, following which no further updates to systems are allowed. Shortly after releasing the test data (without labels or other solutions), the participating teams are to submit their outputs from the frozen systems to the task organizers, who are to evaluate these results and report the resulting benchmarks to the community.

Instead of continuing our replication track from 2016 and 2017 [18,19], we recommended interested teams participate to *ClEf/Ntcir/Trec REproducibility* (CENTRE)[3]. This CENTRE at CLEF 2018 evaluation lab ran a joint CLEF, *NII Testbeds and Community for Information access Research* (NTCIR), and TREC task on challenging participants to study the replicability of selected methods on the same experimental collections as its Task 1; study the reproducibility of selected methods on the different experimental collections as its Task 2; and study the re-reproducibility by using the components developed in aforementioned two tasks and made available by the other participants to replicate/reproduce their results [2]. The CLEF eHealth replication tracks 2016 and 2017 [18,19] gave our participating teams the opportunity to submit their processing methods to organizers, who then attempted to replicate the runs submitted by participants. Three and five participating teams of the CLEF eHealth 2016 Task 2 and the CLEF eHealth 2017 Task 1, respectively, took this opportunity. The teams submitted a total of seven and 22 methods to replication tracks 2016 and 2017, respectively. Both in 2016 and 2017, the organizers were able to achieve a perfect replication, but in some cases, this was only after contacting the submitting team for some further technical clarification on system requirements, installation procedure, and practical use. We were delighted to observe

[3] http://www.centre-eval.org/ (last accessed on 7 June 2018).

an overall improvement in method documentation as an outcome of running the track twice.

The annual CLEF eHealth workhops and evaluation labs have matured and established their presence in 2012–2018. 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 [7,8,13,14,26]. 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[4].

Acknowledgements. The CLEF eHealth 2018 evaluation lab has been supported in part by (in alphabetical order) the ANU, the CLEF Initiative, the Data61/CSIRO, and the French National Research Agency (ANR), under grant CABeRneT ANR-13-JS02-0009-01. 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, SIGIR 2016, pp. 75–84. ACM, New York (2016). https://doi.org/10.1145/2911451.2911510
2. Ferro, N., Maistro, M., Sakai, T., Soboroff, I.: Overview of CENTRE @ CLEF 2018. In: Ferro, N., et al. (eds.) CLEF 2018. LNCS, vol. 11018, pp. 239–246 (2018)
3. Goeuriot, L., et al.: D7.3 meta-analysis of the second phase of empirical and user-centered evaluations. Technical report, Khresmoi Project, August 2014
4. Goeuriot, L., et al.: 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)
5. Goeuriot, L., et al.: An analysis of evaluation campaigns in ad-hoc medical information retrieval: CLEF eHealth 2013 and 2014. Inf. Retr. J. **21**, 1–34 (2018)
6. Goeuriot, L., et al.: 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)
7. Goeuriot, L., et al.: Overview of the CLEF eHealth evaluation lab 2015. In: Mothe, J., et al. (eds.) CLEF 2015. LNCS, vol. 9283, pp. 429–443. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24027-5_44
8. Goeuriot, L., et al.: CLEF 2017 eHealth evaluation lab overview. In: Jones, G.L.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 291–303. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_26
9. Jimmy, Zuccon, G., Demartini, G.: On the volatility of commercial search engines and its impact on information retrieval research. In: SIGIR 2018 (2018)

[4] https://sites.google.com/view/clef-ehealth-2018/home (last accessed on 7 June 2018).

10. 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)
11. 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)
12. 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)
13. 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
14. 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
15. Leeflang, M.M., Deeks, J.J., Takwoingi, Y., Macaskill, P.: Cochrane diagnostic test accuracy reviews. Syst. Rev. **2**(1), 82 (2013)
16. Lipani, A., Palotti, J., Lupu, M., Piroi, F., Zuccon, G., Hanbury, A.: Fixed-cost pooling strategies based on IR evaluation measures. In: Jose, J.M., et al. (eds.) ECIR 2017. LNCS, vol. 10193, pp. 357–368. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-56608-5_28
17. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. **27**(1), 2:1–2:27 (2008). https://doi.org/10.1145/1416950.1416952
18. Névéol, A., et al.: CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org) (2017). ISSN 1613–0073. http://ceur-ws.org/Vol-1866/
19. Névéol, A., et al.: 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) (2016). ISSN 1613–0073. http://ceur-ws.org/Vol-1609/
20. Névéol, A., et al.: 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)
21. Névéol, A., et al.: 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)
22. Palotti, J., et al.: CLEF eHealth evaluation lab 2015, task 2: retrieving information about medical symptoms. In: CLEF 2015 Online Working Notes. CEUR-WS (2015)
23. 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)
24. Park, L.A., Zhang, Y.: On the distribution of user persistence for rank-biased precision. In: Proceedings of the 12th Australasian Document Computing Symposium, pp. 17–24 (2007)
25. Suominen, H.: In: Forner, P., Karlgren, J., Womser-Hacker, C., Ferro, N. (eds.) CLEF 2012 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org) (2012). ISSN 1613–0073. http://ceur-ws.org/Vol-1178/

26. 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
27. 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