Overview of the CLEF eHealth Evaluation Lab 2020

Lorraine Goeuriot[1](✉), Hanna Suominen[2,3,4], Liadh Kelly[5], Antonio Miranda-Escalada[6], Martin Krallinger[6], Zhengyang Liu[2], Gabriella Pasi[7], Gabriela Gonzalez Saez[1], Marco Viviani[7], and Chenchen Xu[2,3]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
Lorraine.Goeuriot@imag.fr,
gabriela-nicole.gonzalez-saez@univ-grenoble-alpes.fr
[2] The Australian National University,
Canberra, ACT, Australia
{hanna.suominen,zhengyang.liu,chenchen.xu}@anu.edu.au
[3] Data61/Commonwealth Scientific and Industrial Research Organisation,
Canberra, ACT, Australia
[4] University of Turku, Turku, Finland
[5] Maynooth University, Maynooth, Ireland
liadh.kelly@mu.ie
[6] Barcelona Supercomputing Center (BSC), Barcelona, Spain
{antonio.miranda,martin.krallinger}@bsc.es
[7] Department of Informatics, Systems, and Communication,
University of Milano-Bicocca, Milan, Italy
{gabriella.pasi,marco.viviani}@unimib.it

Abstract. In this paper, we provide an overview of the eight annual edition of the Conference and Labs of the Evaluation Forum (CLEF) eHealth evaluation lab. The Conference and Labs of the Evaluation Forum (CLEF) eHealth 2020 continues our development of evaluation tasks and resources since 2012 to address laypeople's difficulties to retrieve and digest valid and relevant information in their preferred language to make health-centred decisions. This year's lab advertised two tasks. Task 1 on Information Extraction (IE) was new and focused on automatic clinical coding of diagnosis and procedure the tenth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD10) codes as well as finding the corresponding evidence text snippets for clinical case documents in Spanish. Task 2 on Information Retrieval (IR) was a novel extension of the most popular and established task in the Conference and Labs of the Evaluation Forum (CLEF) eHealth on Consumer Health Search (CHS). In total 55 submissions were made to these tasks. Herein, we describe the resources created

With equal contribution, LG, HS & LK co-chaired the lab. The leaders of Task 1 were AM-E and MK. The leaders of Task 2 were LG and HS, with LK, ZL, GP, GGS, MV, and CX as co-organizers and contributors to the evaluation conceptualization, dataset creation, assessments, and measurements.

A. Arampatzis et al. (Eds.): CLEF 2020, LNCS 12260, pp. 255–271, 2020.
https://doi.org/10.1007/978-3-030-58219-7_19

for the two tasks and evaluation methodology adopted. We also summarize lab submissions and results. As in previous years, the organizers have made data and tools associated with the lab tasks available for future research and development. The ongoing substantial community interest in the tasks and their resources has led to the Conference and Labs of the Evaluation Forum (CLEF) eHealth maturing as a primary venue for all interdisciplinary actors of the ecosystem for producing, processing, and consuming electronic health information.

Keywords: eHealth · Evaluation · Health records · Medical informatics · Information extraction · Information storage and retrieval · Speech recognition · Test-set generation

1 Introduction

Easy-to-understand *Electronic Health Records* (EHRs) can contribute to patients' right—and, if applicable, also their home-based carers or other next-of-kins' right—to be informed about their health and health care. The requirement to ensure that patients can understand their official, privacy-sensitive health information in their own EHR is stipulated by policies and laws [21]. Improving patients' ability to access and digest this content could mean paraphrasing the EHR-text, enriching it with hyperlinks to term definitions, care guidelines, and further supportive information on patient-friendly and reliable websites, helping them to discover good search queries to retrieve more contents, allowing not only text but also speech as a query modality, enabling search in multiple languages, and developing methods for such reading aids to release health care workers' time from EHR-writing to, for example, longer patient-education discussions [39,41].

The *Conference and Labs of the Evaluation Forum* (CLEF) and other information access conferences have organized evaluation labs on related *Electronic Health* (eHealth) *Information Extraction* (IE), *Information Management* (IM), and *Information Retrieval* (IR) tasks for approximately 20 years. Yet they have predominantly targeted the health care experts' information needs only [2,3,12]. A rare exception is the annual *CLEF eHealth Evaluation-lab and Lab-workshop Series* from 2012 to 2020 [9,10,15–17,38,42,44]. In 2012, the first scientific CLEF workshop took place, with an aim of establishing an evaluation campaign, and from 2013 to 2020, this annual workshop has been supplemented with a lead-up evaluation lab, consisting of up to three shared tasks each year. Although the tasks have been centered around the patients and their families' needs in accessing and understanding eHealth information, additional use cases were also addressed in 2015–2019, for example, *Automatic Speech Recognition* (ASR) and IE to aid clinicians in IM.

In 2020, CLEF eHealth advertised two tasks. Task 1 on IE was new and focused on clinical coding of terms or evidence for assigning diagnosis or procedure codes to clinical textual data in Spanish. Task 2 on IR included a traditional adhoc task, as well as a novel extension of the adhoc task with spoken queries on

Consumer Health Search (CHS). Further details on the previous task/problem specifications and data and methods releases of these tasks are available in [39] and [17].

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 2020 evaluation lab. Finally, we compare with prior editions of CLEF eHealth and conclude the paper.

2 Materials and Methods

In this section, we describe the materials and methods used in the two tasks of the CLEF eHealth evaluation lab 2020. 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. For the 2020 Task 1 (abbreviated as CodiEsp; promoted by the Plan de Impulso de las Tecnologías del Lenguaje - Plan TL, https://www.plantl.gob.es) we used the SPACCC corpus of Spanish clinical case documents [1,13], a collection of 1,000 carefully selected clinical cases resembling EHRs classified manually using the MyMiner File Labelling tool [35] by a practicing physician with assistance of a clinical documentalist. This dataset was already exploited previously for other shared tasks related to the automatic detection of drugs, chemical compounds and genes (PharmaCoNER Track, [1]) and partially for the detection and resolution of medical abbreviations (BARR2, [13]). Overall, this corpus contains a total of 16,504 sentences and 396,988 tokens, with an average of 396.99 tokens per clinical case, thus these records are considerably longer than the data used by past CLEF clinical coding tasks employing death certificates [24–26] and non-technical summaries of animal experimentation [27].[1]

As this corpus includes records from a variety of clinical disciplines, such as oncology, cardiology, ophthalmology, urology or infectious diseases it covers a great diversity of clinical fields, increasing the complexity for natural language processing tasks.

Mentions of diagnostics and medical procedures evidence text snippets were annotated manually and mapped to the *tenth revision of the International Statistical Classification of Diseases and Related Health Problems* (ICD10) codes by experts in clinical coding to generate a Gold Standard corpus (annotation information is depicted in Sect. 2.2). Thereafter, we randomly generated three

[1] The CodiEsp corpus, together with the other generated resources are available at the *Medical Natural Language Processing* (NLP) Zenodo community, https://zenodo.org/communities/medicalnlp/ and at the shared task webpage, https://temu.bsc.es/codiesp/.

non-overlapping subsets: training set (500 documents), development and test set (250 documents each).

To facilitate comparison to systems working with data in English, and explore the use of machine translation technologies to extend or complement traditional corpus construction approaches, we also provided participants an automatically translated version of our corpus into English (CodiEsp *Machine Translation* (MT) corpus). Therefore we constructed a machine translation approach (English-Spanish) adapted to the language characteristics of the medical domain [37].

To use this task setting to extend the initial CodiEsp corpus (Gold Standard manual annotations) by generating a silver standard consisting of automatic annotations generated by participating teams, similar to the CALBC initiative [33], we added to the test set an additional collection of 2,751 documents [22].

It is noteworthy to point out that a corpus of only 1,000 documents for a complex clinical coding task with thousands of possible codes or class labels is rather small to fully exploit the predictive power of more advanced machine learning approaches.

To overcome this issue, we generated two additional data collections, exploiting existing mappings between clinical coding terms from the *tenth revision of the International Statistical Classification of Diseases and Related Health Problems* (ICD10) to *Medical Subject Headings* (MeSH) terms using the *Unified Medical Language System* (UMLS) Metathesaurus. Moreover, in turn most *Medical Subject Headings* (MeSH) terms do have a corresponding DeCS code. Thus by using the mapping chain [DeCS →*Medical Subject Headings* (MeSH) →*Unified Medical Language System* (UMLS) →the *tenth revision of the International Statistical Classification of Diseases and Related Health Problems* (ICD10)] we could generate a collection of medical literature manually indexed with either DeCS or *Medical Subject Headings* (MeSH) and index them with their corresponding the *tenth revision of the International Statistical Classification of Diseases and Related Health Problems* (ICD10) codes resulting in the CodiEsp-abstracts corpus. It is composed of 176,294 Spanish medical abstracts indexed with the *tenth revision of the International Statistical Classification of Diseases and Related Health Problems* (ICD10) codes.

Task 2. The 2018 CLEF eHealth Consumer Health Search document collection was used in this year's IR challenge. As detailed in [14], this collection consists of web pages acquired from the CommonCrawl. An initial list of websites was identified for acquisition. The list was built by submitting queries on the 2018/2020 topics 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. See [11] for full details on the Task 2 dataset.

2.2 Human Annotations, Queries, and Relevance Assessments

Task 1. Due to the complexity and practical importance of clinical coding the CodiEsp corpus was generated by a team of professional clinical coding experts. In addition to assigning clinical codes they also had to label the textual evidence supporting the code assignment.

To assure quality and to determine the difficulty of this task, the annotation process followed an iterative annotation team training exercise until a satisfactory *Inter-Annotator Agreement* (IAA) was reached. Finally, a set of 50 records were double annotated (blinded) by two different expert annotators, reaching a pairwise agreement of 80.5% on the annotations of the evidence text spans and of 88.6% for the assignment of documents to diagnostic codes and 88.9% for procedure codes.

We released the plain text documents together with a tab-separated file with the annotation information similar to the format employed in past CLEF clinical coding tasks [27].

The CodiEsp corpus covers 3,427 unique ICD-10 codes corresponding to a total of 18,435 manual document-code annotations. The most common code is r52, corresponding to "unspecified pain"; which is repeated 361 times across the entire corpus. 1,830 codes appear more than once, among which 346 codes appear more than 10 times. A large amount of infrequent codes poses an extra challenge for CodiEsp participants, which some of them have surpassed using the additional corpus, CodiEsp-abstracts.

Task 2. Historically the CLEF eHealth IR task has released text queries representative of layperson medical information needs in various scenarios. In recent years query variations issued by multiple laypeople for the same information need have been offered. In this year's task, we extended this to spoken queries. These spoken queries were generated by 6 individuals using the information needs derived for the 2018 challenge. We also provided textual transcripts of these spoken queries and automatic speech-to-text translations. The topics for the adhoc subtask were similar to 2018 CHS task topics: 50 queries, which were issued by the general public to the HON (*Health on the Net*) search service. These queries were manually selected by a domain expert from a sample of raw queries collected over a period of 6 months to be representative of the type of queries posed to the search engine. Queries were not preprocessed, for example any spelling mistakes that may be present have not been removed. All the queries from the adhoc task have been recorded with several users for the subtask on Spoken queries retrieval. A transcription of these audio files was also provided, using ESPNET, Librispeech, CommonVoice and Google API (with three models). Spoken queries could be downloaded from a secured server, with an agreement signed by the participating team.

The relevance assessment has been conducted on three relevance dimensions: topicality, understandability and credibility. Topicality is a classical relevance dimension ensuring that the document and the query are on the same topic and the document answers the query. Understandability is an estimation of whether the document is understandable by a patient. Topicality and understandability have been used as relevance dimensions in the CHS task of CLEF eHealth for several years. This year, we introduced a novel dimension, i.e., *credibility*, which is as a perceived quality of the information receiver. It is composed of *multiple dimensions* that have to be considered and evaluated together in the process of information credibility assessment [4,36]. In the health-related context, the multiple dimensions that have to be considered when evaluating information credibility are related to the *source* that disseminate a content, the characteristics related to the *message* diffused, and *social aspects* if the information is disseminated through virtual communities [45]. Therefore, the assessors were asked to evaluate the above-mentioned multiple aspects by considering, at the same time, any information available about the trustworthiness of the source of the health-related information [20] (the fact that information comes from a Web site with a good or bad *reputation*, or the level of *expertise* of an individual answering on a blog or a question-answering system, etc.), the *syntactic/semantic characteristics* of the content [5] (in terms of completeness, language register, style, etc.), and any information emerging from social interactions [32] (the fact circle of social relationships of the author of a content is reliable or not, the fact that the author is involved in many discussions, etc.). All the dimensions were considered on a 3-levels scale:

- *not relevant/understandable/credible*
- *somewhat relevant/understandable/credible*
- *highly relevant/understandable/credible*
- We added a 4th option for credibility for assessors uncertainty: *I am not able to judge.*

Relevance assessments are currently in progress.

Similar to the 2016, 2017 and 2018 pools, we created the pool using the RBP-based Method A (Summing contributions) by Moffat et al. [23], 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 [31]). 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 [19], as it is the case for this task. As the topics were similar, the pool is an extension of 2018's pool.

2.3 Evaluation Methods

Task 1. Task 1 was composed of three distinct subtasks: CodiEsp-Diagnostic, CodiEsp-Procedure, and CodiEsp-Explainability. Participants of the CodiEsp-Diagnostic and CodiEsp-Procedure tracks predicted the ICD-10 codes for the 250 documents contained in the test set. Predictions were compared or assessed against manually assigned annotations. CodiEsp-Explainability participants had to predict not only the codes but also the corresponding textual evidence snippets to enable human interpretation or validation of automatic assignments. For the

CodiEsp-Diagnostic and CodiEsp-Procedure subtasks, the codes assigned to each document had to be ranked, providing high confidence codes on the top of the list. Thus more relevance was given to predictions for which the system was more confident. The main metric for these two subtasks was Mean Average Precision (MAP). MAP is a widely established metric in ranking problems and was used by other challenges like TREC. It stands for Mean Average Precision, where the Average Precision represents the average precision of a document at every position in the ranked codes. That is, precision is computed considering only the first ranked code; then, it is computed considering the first two codes, and so on. Finally, precision values are averaged over the number of codes in the gold standard (the relevant number of codes).

For completeness, error analysis, and comparison reasons, other metrics are computed for these two subtasks: MAP@k (MAP taking into account just the first k results), f-score, precision, and recall.

Since the scope of the explainability subtask is different and more challenging, participants were evaluated with f-score, precision, and recall.

Task 2. For Subtasks 1 and 2, participants could submit up to 4 runs in TREC format. Evaluation measures are NDCG@10, BPref and RBP. Metrics such as uRBP will be used to capture various relevance dimensions.

3 Results

CLEF eHealth tasks offered every year in 2013–2020 have brought together researchers working on health information access topics. It has provided them with data and computational resources to work with and validate their outcomes. These contributions of the lab have accelerated pathways from scientific ideas through influencing research and development to societal impact. Targeted use scenarios for the designed, developed, and evaluated technologies have included easing patients, their families, clinical staff, health scientists, and health care policy makers in accessing and understanding health information. Its niche is addressing health information needs of laypeople (including, but not limited to, patients, their families, clinical staff, health scientists, and health care policy makers)—and not health care experts only—in a range of languages—in retrieving and digesting valid and relevant eHealth information to make health-centered decisions [2,3,12,39,40].

By 2020, the CLEF eHealth evaluation lab has matured as a popular primary venue for all interdisciplinary actors of the ecosystem for producing, processing, and consuming eHealth information. In 2013, 2014, 2015, 2016, 2017, 2018, 2019, and 2020 as many as 170, 220, 100, 116, 67, 70, 67, and 57 teams have registered their expression of interest in the CLEF eHealth tasks, respectively, and the number of teams proceeding to the task submission stage has been 53, 24, 20, 20, 32, 28, 9, and 55 respectively [9,10,15–17,42,44].[2] In 2020, 51 and 24 teams registered to CLEF eHealth Task 1 and Task 2, respectively; 18 teams expressed their interest in this way to both offered tasks. Of the 55 CLEF eHealth submissions in 2020, 22 targeted the CodiEspD Diagnostic subtask of Task 1, 17 the CodiEspP Procedure subtask of Task 1, and 8 the CodiEspX Explainability subtask of Task 1. Among five submission to the 2020 CLEF eHealth Task 2, the ad hoc IR subtask was the most popular with its three submissions; the subtasks that used transcriptions of the spoken queries and the original audio files received one submission each.

Next, more details about the task outcomes are presented. See [22] and [11] for further details.

3.1 Task 1

51 teams registered for Task 1 (CodiEsp), out of which 22 submitted predictions for at least one of the three subtracks. We allowed a total of 5 runs for each sub-track, so that teams could explore different approaches. 47 submissions were made in total, 22 for subtask CodiEsp-Diagnostic, 17 for CodiEsp-Procedure, and 8 for CodiEsp-Explainability. The number of submitted runs were: 78 for CodiEsp-Diagnostic, 64 for CodiEsp-Procedure, and 25 for CodiEsp-Explainability. In total, 167 clinical coding systems were created in the context of Task 1.

From the 22 participant teams, 3 reported being a commercial organization. Despite the fact that the used data was in Spanish, the participation was global covering teams not only from Spanish-speaking countries (Spain and Argentina) but also from India, Italy, Germany, United States, Japan, France, Belgium, Turkey, and the UK.

All best-performing teams obtained higher results than the baseline. In CodiEsp-Diagnostic, the best Mean Average Precision result has been 0.593, obtained by the team IXA-AAA. In the CodiEsp-Procedure subtask, team IAM obtained 0.493 MAP, the best result. For comparison purposes with past clinical coding shared tasks, we also provide the best results in terms of f1-score, precision, and recall. In CodiEsp-Diagnostic, the highest achieved f1-score was 0.687; the highest precision was 0.866, and the highest recall was 0.897. In CodiEsp-Procedure, they were 0.522, 0.833, and 0.825, respectively. Finally, in CodiEsp-Explainability, two teams (FLE and IAM) achieved 0.611 f1-score, 0.75 was the top precision, and 0.562 the top recall. In the three subtasks, teams that developed the highest-performing systems were closely followed by others.

3.2 Task 2

The 2020 CLEF eHealth Task 2 attracted five submissions (Table 1). Its ad hoc IR subtask was the most popular (three submissions). Two of these teams also

[2] "Expressing an interest" for a CLEF task consists of filling in a form on the CLEF conference website with contact information, and tick boxes corresponding to the labs of interest.

submitted to the subtasks based on spoken queries. Specifically, the subtask that used transcriptions of the spoken queries had one submission and the subtask where the original audio files were processed had one submission. The submitting teams were from Australia, France, and Italy and had 4, 1, and 6 team members, respectively. They were all from academia and each team had members from a single organization.

Although these submission numbers were considerably smaller than in the seven previous years of running the CHS task [39–41], the organizers were pleased with this newly introduced task, with its novel spoken queries element attracting interest and submissions.

Table 1. Descriptive statistics about teams that submitted to the CLEF eHealth 2020 Task 2

Subtasks	No. of coauthors	Authors' affliction	Affiliation country
Ad Hoc search & spoken queries using transcriptions	1	1 university	Italy
Ad Hoc search & spoken queries using audio files	6	1 university	France
Ad Hoc search	4	1 university	Australia

The Italian submission to the Ad Hoc Search and Spoken Queries Using Transcription subtasks was by Associate Professor Giorgio Maria Di Nunzio from the Information Management System (IMS) Group of the University of Padua. His submission to the former task included BM25 of the original query; Reciprocal Rank fusion with BM25, *Query Language Model* (QLM), and *Divergence from Randomness* (DFR) approaches. Reciprocal Rank fusion with BM25, QLM, and DFR approaches using pseudo relevance feedback with 10 documents and 10 terms (the query weight of 0.5); and Reciprocal rank fusion with BM25 run on manual variants of the query. His submission to the latter task included the Reciprocal Rank fusion with BM25; Reciprocal Rank fusion with BM25 using pseudo relevance feedback with 10 documents and 10 terms (the query weight of 0.5); Reciprocal Rank fusion of BM25 with all transcriptions; and Reciprocal Rank fusion of BM25 with all transcripts using pseudo relevance feedback with 10 documents and 10 terms (the query weight of 0.5).

The French team was formed by Dr Philippe Mulhem, Aidan Mannion, Gabriela Gonzalez Saez, Associate Professor Didier Schwab, and Jibril Frej from the Laboratoire d'Informatique de Grenoble of the Univ. Grenoble Alpes. Their team name was LIG-Health. To the Ad Hoc Search task, they submitted runs using Terrier BM25 as a baseline, and explored various expansion methods using UMLS, using the Consumer Health Vocabulary, expansion using Fast Text; and Terrier BM25 with RF (bose-Einstein) weighted expansion. For the

Spoken Queries they used various transcriptions on the same models, opting for the best performing ones based on 2018 qrels. They submitted merged runs for each query.

The Australian team—called SandiDoc from the Our Health In Our Hands (OHIOH) Big Data program, Research School of Computer Science, College of Engineering and Computer Science, The Australian National University had Sandaru Seneviratne, Dr Eleni Daskalaki, Dr Artem Lenskiy, and Dr Zakir Hossain as its members—took part in the Ad Hoc Search task with a method founded on $TF \times IDF$ scoring. First, they pre-processed both the queries and the dataset. Then, they obtained TF×ID scores for the queries and used these TF × ID scores to obtain the most similar documents for the queries. Finally, they supplemented this method by working on the clefehealth2018_B dataset using the medical skipgram word embeddings (vectors_medtrack_skipgram_s500_w5_neg20_hs0_sam1e-4_iter5) provided. To represent the documents and queries, they used the average word vector representations as well as the average of minimum and maximum vector representations of the document or query. In documents, these representations were obtained using the 100 most frequent words in a document. For each of these two representations, they calculated the similarity among documents and queries using the cosine measure to obtain the final results for the task. The aim was to experiment with different vector representations for text.

In addition to these participants' methods, we as the organizers developed baseline methods that were based on the renown OKapi BM25 but now with REINFORCE based query expansion. This baseline method had the following two phases: First, the initial query was enriched with a query expansion model, which was pre-trained on general corpora and then used to retrieve documents by reusing the commonly-used BM25 algorithm [34]. Second, in the query expansion phase, the system was optimized in an reinforcement learning paradigm as proposed in [28]. Given an original query, the system performed trials of generating new queries and rewarded them by matching the documents retrieved from these queries against the ground truth ranking. The context words in the newly retrieved documents also contributed to the construction of queries for the next iteration in order to ensure enough data sources for the learning process. This baseline adopted the pre-trained model optimized on the TREC-CAR, Jeopardy, and MSA datasets [28]. Once the query was expanded to a few related candidates, they were fed to a general implementation of BM25 algorithm to retrieve the final set of documents.

The intuition behind this query expansion was that a layperson may lack the professional knowledge to accurately describe medical terms; differently to the rigorous wording in the medical documents to be retrieved, a layperson's input query usually contains inexact and long descriptions. Thus, query expansion was applied to automatically rewrite the query in a way that increases the probability of matching more candidates. In this baseline method, we employed the REINFORCE algorithm introduced in [46]. Given an original query q_0, it retrieved some ranked documents D_0 and from where new candidate query q'_0 was constructed. The new query q'_0 was fed back into the retrieval system to produce ranked documents D'_0. This process of documents retrieval and construction of new query was iterated to create training examples $\{(q'_0, D'_0), (q'_1, D'_1), \ldots\}$. At each step, the operations adopted to reformulate the new query were recorded as the actions. The retrieved documents D'_k were then compared against the ground truth ranking result to calculate the reward for this new query and so were the actions to generate it. The process was optimized in a reinforcement learning paradigm to learn a system that could generate a series of candidate queries from an input one. In particular, the stochastic objective to optimize was:

$$C_a = (R - \bar{R}) \sum_{t \in T} - \log P\left(t \mid q_0\right),$$

where R and $\bar{R}$ are the reward from the new query and baseline reward, and $t \in T$ are words from the new query.

The relevance assessments are being collected at the time of writing of this paper. See the Task 2 overview paper for further details and the results of the evaluation [11].

4 Comparison with Prior CLEF eHealth Work

Since its inception the CLEF eHealth lab series has offered IE and IR shared challenges. In particular, IE challenges related to ICD-10 coding started in 2016, and query driven IR challenges started in 2013 with the commencement of the lab.

4.1 Information Extraction

CLEF eHealth 2016 evaluation lab [24] challenged participants to assign ICD-10 codes to death certificates in French. The corpus contained a collection of sentences extracted from 27,850 death certificates. The coding of these documents is relevant to guide public health policies. There were 5 participant teams, teams achieved 0.719 f1-score on average and the best run reached 0.848 f1-score. In terms of precision, the best run reached 0.813 and in terms of recall, 0.890.

On CLEF eHealth 2017 Multilingual Information Extraction task [25], participants again had to assign ICD-10 codes to sentences extracted from death certificates. In this case, both in English and French. The corpus contained a collection of sentences extracted from 31,690 French death certificates and 6,665 English death certificates. There were 10 competing systems for the English subtask and 9 for the French one. The highest performance in French was 0.867 f1-score. The highest precision was 0.881 and highest recall 0.875.

The same setting was replicated on CLEF eHealth 2018 Multilingual Information Extraction task [26]. However, death certificates in Hungarian and Italian are added in this edition. There was higher participation, with 14 teams working on French death certificates, 5 on Hungarian, and 6 on Italian. The best systems achieved 0.838 f1-score on French, 0.963 on Hungarian, and 0.952 on Italian.

The best precision scores were 0.835, 0.955 and 0.945. Finally, best recalls were 0.846, 0.97 and 0.96.

CLEF eHealth 2019 Multilingual Information Extraction [27] proposed a novel type of document to code, non-technical summaries of animal experimentation. Coding them is relevant to support the analysis of animal experimentation data. In addition, this year's evaluation lab provided the documents in German language and the goal of the edition was to assign ICD-10 codes to the complete document, not to individual sentences. There were 6 competing teams and the top-performing team achieved an f1-score of 0.80. The best recall was 0.86 and the best precision 0.98.

This year's CLEF eHealth Multilingual Information Extraction task (CodiEsp) introduced a clinical coding challenge in a new language, Spanish; on a new kind of document, clinical case reports; and with a different evaluation metric, Mean Average Precision. It has attracted a higher interest within the community since the number of participants has increased to 22 (Fig. 1).

The CodiEsp corpus was more complex (longer documents covering heterogeneous clinical specialties) than in the shared tasks from 2016, 2017 and 2018; since death certificates are much shorter narratives than clinical case records [18]. In addition, these past shared tasks involved clinical coding of diagnostics; while this year's shared task included a subtask on clinical coding of procedures. Finally, the dataset employed in 2016, 2017 and 2018 was more extensive: for 3,457 unique codes, there were 377,677 code assignments; while our contains 18,435 annotations for 3427 unique codes.

The CodiEsp corpus was also more complex than the corpus of 2019's shared task. Both had a similar number of code assignments, but the latter included 233 distinct codes, while our corpus had 3,427 unique codes. However, the CodiEsp corpus provided the textual evidence supporting the coding decision, which has helped teams building systems not based on document or sentence classification and partially bridges the complexity gap between both datasets.

This increase in complexity may have been one of the reasons for the differences in team performance (see Sect. 3 on results). However, it is noteworthy that team IAM, which employed a similar method in 2018's and in this year's shared tasks obtained comparable f1-scores: 0.666 in the French raw clinical coding and 0.687 in CodiEsp-Diagnostic.

4.2 Information Retrieval

In 2013 and 2014 the focus of the IR task was on evaluating the effectiveness of search engines to support people when searching for information about known conditions, for example, to answer queries like "thrombocytopenia treatment corticosteroids length", with multilingual queries added in the 2014 challenge [6–8]. This task aimed to model the scenario of a patient being discharged from hospital and wanting to seek more information about diagnosed conditions or prescribed treatments.

In 2015 the IR task changed to focus on studying the effectiveness of search engines to support individuals' queries issued for self-diagnosis purposes, and

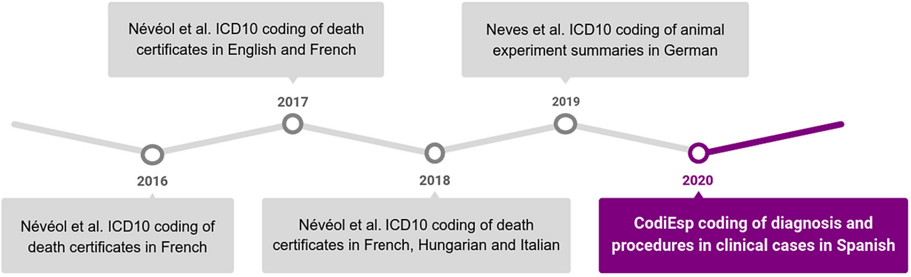
Névéol et al. ICD10 coding of death certificates in English and French
Neves et al. ICD10 coding of animal experiment summaries in German
2017
2019
2016
2018
2020
Névéol et al. ICD10 coding of death certificates in French
Névéol et al. ICD10 coding of death certificates in French, Hungarian and Italian
CodiEsp coding of diagnosis and procedures in clinical cases in Spanish

Fig. 1. Clinical coding CLEF eHealth shared tasks.

again offered a multilingual queries challenge [29]. In addition, we began adding personalization elements to the challenge on an incremental basis by assessing the readability of information and taking this into account in the evaluation framework.

This individualized IR approach was continued in the 2016 and 2017 labs [30,47] and we also introduced gradual shifts from an ad-hoc search paradigm (that of a single query and a single document ranking) to a session based search paradigm. Along these lines we also revised how relevance is measured for evaluation purposes, taking into account instead whole-of-session usefulness.

In 2018 [14] we continued this evolution, and introduced query intent elements. 7 teams participated in this challenge. The IR task did not run in 2019.

This year's challenge built on the 2018 challenge by introducing a new spoken query element, whereby participants had the additional optional challenge of retrieving using speech-to-text translations of the queries. This challenge used the same document collection as that used in 2018 and also the same topics, for which new spoken queries were generated by 6 individuals accounting for 6 query variants. Text transcripts of the queries were also available. The primary new element then of this year's challenge was the provision of spoken queries and speech-to-text translations of these queries.

The CHS task has been exploring for several years health documents relevance and its dimensions. This has been a great success and has led to the creation of systems better suited to the patients. The introduction of the credibility in the dimension this year is another step towards better and safer health information online.

Given 5 teams participated in this year's challenge relative to the 7 in the 2018 IR challenge, one might conjecture that the optional use of spoken queries was off putting for potential teams. However, both 2018 and 2020 participant figures are down on the earlier years of the IR challenge, where in excess of 10 teams participated each year. Earlier cycles of the IR challenge adopted a simpler ad-hoc IR challenge approach and used simpler IR metrics. We have found that as the task increased in complexity and further options for participation in the form of subtasks have been added that the number of participating teams has decreased. That being said, we find much use of the datasets post CLEF [39].

5 Conclusions

This paper provided an overview of the CLEF eHealth 2020 evaluation lab. The inaugural CLEF eHealth workshop took place in 2012 with an aim of establishing an evaluation lab [38]. This ambition was realised in 2013, with annual CLEF eHealth evaluation labs and workshops organized every year since 2013 [9,10,15–17,43,44]. In 2020, it ran an IE task in Spanish and IR task in English.

During these past nine years, the CLEF eHealth series has continuously offered carefully designed and well resourced evaluation tasks to the research and development community. This contribution includes, but is not limited to, the creation and dissemination of speech and text analytics resources such as problem/task specifications, test collections, annotations, assessments, annotation/assessment methods, processing methods, evaluation methods, and evaluation benchmarks in understanding, accessing, and authoring health information in a multilingual setting.

Given the significance of the CLEF eHealth community, tasks, and resources over the years, our aim is to keep the tasks going in years to come. Our releases so far can be found on our CLEF eHealth website[3].

Acknowledgements. We gratefully acknowledge the contribution of the people and organizations involved in CLEF eHealth in 2012–2020 as participants or organizers. We thank the CLEF Initiative, Benjamin Lecouteux (Université Grenoble Alpes), João Palotti (Qatar Computing Research Institute), and Guido Zuccon (University of Queensland). We also thank the individuals who generated spoken queries for the IR challenge. We are very grateful to our assessors that helped despite the COVID-19 crisis: Paola Alberti, Vincent Arnone, Nathan Baran, Pierre Barbe, Francesco Bartoli, Nicola Brew-Sam, Angela Calabrese, Sabrina Caldwell, Daniele Cavalieri, Madhur Chhabra, Luca Cuffaro, Yerbolat Dalabayev, Emine Darici, Marco Di Sarno, Mauro Guglielmo, Weiwei Hou, Yidong Huang, Zhengyang Liu, Federico Moretti, Marie Revet, Paritosh Sharma, Haozhan Sun, Christophe Zeinaty. The lab has been supported in part by (in alphabetical order) The Australian National University, College of Engineering and Computer Science, Research School of Computer Science; and the CLEF Initiative. We acknowledge the Encargo of Plan TL (SEAD) to CNIO and BSC for funding, and the scientific committee for their valuable comments and guidance.

References

1. Agirre, A.G., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., Krallinger, M.: Pharmaconer: pharmacological substances, compounds and proteins named entity recognition track. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, pp. 1–10 (2019)
2. Demner-Fushman, D., Elhadad, N.: Aspiring to unintended consequences of natural language processing: a review of recent developments in clinical and consumer-generated text processing. Yearb. Med. Inform. **1**, 224–233 (2016)

[3] https://clefehealth.imag.fr/ (last accessed on 19 June 2020).

3. Filannino, M., Uzuner, Ö.: Advancing the state of the art in clinical natural language processing through shared tasks. Yearb. Med. Inform. **27**(01), 184–192 (2018)
4. Fogg, B.J., Tseng, H.: The elements of computer credibility. In: Proceedings of SIGCHI (1999)
5. Fontanarava, J., Pasi, G., Viviani, M.: Feature analysis for fake review detection through supervised classification. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 658–666. IEEE (2017)
6. 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)
7. Goeuriot, L., et al.: An analysis of evaluation campaigns in ad-hoc medical information retrieval: CLEF eHealth 2013 and 2014. Inf. Retriev. J. **21**(6), 507–540 (2018). https://doi.org/10.1007/s10791-018-9331-4
8. 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, England (2014)
9. 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
10. Goeuriot, L., et al.: CLEF 2017 eHealth evaluation lab overview. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 291–303. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_26
11. Goeuriot, L., et al.: Overview of the CLEF eHealth 2020 task 2: consumer health search with ad hoc and spoken queries. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2020)
12. Huang, C.C., Lu, Z.: Community challenges in biomedical text mining over 10 years: Success, failure and the future. Briefings Bioinform. **17**(1), 132–144 (2016)
13. Intxaurrondo, A., et al.: Finding mentions of abbreviations and their definitions in spanish clinical cases: the barr2 shared task evaluation results. In: IberEval@ SEPLN, pp. 280–289 (2018)
14. Jimmy, J., 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)
15. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth evaluation lab 2016. In: Fuhr, N., Quaresma, P., Gonçalves, T., Larsen, B., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 255–266. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_24
16. 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
17. Kelly, L., et al.: Overview of the CLEF eHealth evaluation lab 2019. In: Crestani, F., et al. (eds.) CLEF 2019. LNCS, vol. 11696, pp. 322–339. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-28577-7_26
18. Lavergne, T., Névéol, A., Robert, A., Grouin, C., Rey, G., Zweigenbaum, P.: A dataset for ICD-10 coding of death certificates: creation and usage. In: Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016), pp. 60–69. The COLING 2016 Organizing Committee, Osaka, Japan, December 2016. https://www.aclweb.org/anthology/W16-5107

19. 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
20. Livraga, G., Viviani, M.: Data confidentiality and information credibility in on-line ecosystems. In: Proceedings of the 11th International Conference on Management of Digital EcoSystems, pp. 191–198 (2019)
21. McAllister, M., Dunn, G., Payne, K., Davies, L., Todd, C.: Patient empowerment: the need to consider it as a measurable patient-reported outcome for chronic conditions. BMC Health Serv. Res. **12**, 157 (2012)
22. Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estapé, J., Krallinger, M.: Overview of automatic clinical coding: annotations, guidelines, and solutions for non-English clinical cases at codiesp track of CLEF eHealth 2020. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2020)
23. 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
24. 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/
25. 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)
26. 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)
27. Neves, M., et al.: Overview of task 1 in CLEF eHealth 2019: indexing German non-technical summaries of animal experiments. In: CLEF 2019 Online Working Notes. CEUR-WS (2019)
28. Nogueira, R., Cho, K.: Task-oriented query reformulation with reinforcement learning. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2017). https://doi.org/10.18653/v1/d17-1061
29. 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)
30. 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)
31. 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)
32. Pasi, G., Viviani, M.: Information credibility in the social web: Contexts, approaches, and open issues. arXiv preprint arXiv:2001.09473 (2020)
33. Rebholz-Schuhmann, D., et al.: CALBC silver standard corpus. J. bioinform. Comput. Biol. **8**(01), 163–179 (2010)
34. Robertson, S.: The probabilistic relevance framework: BM25 and beyond. Found. Trends® Inf. Retriev. **3**(4), 333–389 (2010). https://doi.org/10.1561/1500000019
35. Salgado, D., et al.: MyMiner: a web application for computer-assisted biocuration and text annotation. Bioinformatics **28**(17), 2285–2287 (2012)

36. Self, C.C.: Credibility. In: An Integrated Approach to Communication Theory and Research, pp. 449–470. Routledge (2014)
37. Soares, F., Krallinger, M.: BSC participation in the WMT translation of biomedical abstracts. In: Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pp. 175–178 (2019)
38. Suominen, H.: CLEFeHealth2012 – The CLEF 2012 workshop on cross-language evaluation of methods, applications, and resources for eHealth document analysis. In: Forner, P., Karlgren, J., Womser-Hacker, C., Ferro, N. (eds.) CLEF 2012 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org) (2012). ISSN 1613–0073, http://ceur-ws.org/Vol-1178/
39. Suominen, H., Kelly, L., Goeuriot, L.: Scholarly influence of the conference and labs of the evaluation forum eHealth Initiative: review and bibliometric study of the 2012 to 2017 outcomes. JMIR Res. Protoc. **7**(7), e10961 (2018). https://doi.org/10.2196/10961
40. Suominen, H., Kelly, L., Goeuriot, L.: The scholarly impact and strategic intent of CLEF eHealth labs from 2012 to 2017. Information Retrieval Evaluation in a Changing World. TIRS, vol. 41, pp. 333–363. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-22948-1_14
41. Suominen, H., Kelly, L., Goeuriot, L., Krallinger, M.: CLEF ehealth evaluation lab 2020. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval, pp. 587–594. Springer International Publishing, Cham (2020)
42. Suominen, H., et al.: Overview of the CLEF eHealth evaluation lab 2018. In: Bellot, P., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 286–301. Springer , Cham (2018). https://doi.org/10.1007/978-3-319-98932-7_26
43. Suominen, H., et al.: Overview of the CLEF ehealth evaluation lab 2018. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 286–301. Springer, Heidelberg (2018)
44. 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
45. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health information–a survey. Wiley Interdisc. Rev.: Data Mining Knowl. Disc. **7**(5), e1209 (2017)
46. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. In: Reinforcement Learning, pp. 5–32. Springer, US (1992). https://doi.org/10.1007/978-1-4615-3618-5_2
47. 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