Overview of the CLEF eHealth Evaluation Lab 2021

Hanna Suominen, Lorraine Goeuriot, Liadh Kelly, Laura Alonso Alemany, Elias Bassani, Nicola Brew-Sam, Viviana Cotik, Darío Filippo, Gabriela González-Sáez, Franco Luque, et al.

Overview of the CLEF eHealth Evaluation Lab 2021

Hanna Suominen[1,2,3], Lorraine Goeuriot[4], Liadh Kelly[5], Laura Alonso Alemany[6], Elias Bassani[7,8], Nicola Brew-Sam[1], Viviana Cotik[9,10], Darío Filippo[11], Gabriela González-Sáez[4], Franco Luque[6,12], Philippe Mulhem[4], Gabriella Pasi[7], Roland Roller[13], Sandaru Seneviratne[1], Rishabh Upadhyay[7], Jorge Vivaldi[14], Marco Viviani[7], and Chenchen Xu[1,2] *

[1] The Australian National University, Canberra, ACT, Australia, firstname.lastname@anu.edu.au
[2] Data61/Commonwealth Scientific and Industrial Research Organisation, Canberra, ACT, Australia
[3] University of Turku, Turku, Finland
[4] Université Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France, firstname.lastname@imag.fr
[5] Maynooth University, Ireland, liadh.kelly@mu.ie
[6] Universidad Nacional de Córdoba, Argentina, {lauraalonsoalemany,francolq}@unc.edu.ar
[7] University of Milano-Bicocca, DISCo, Italy, firstname.lastname@unimib.it
[8] Consorzio per il Trasferimento Tecnologico - C2T, Milan, Italy, elias.bassani@consorzioc2t.it
[9] Departamento de Computación, FCEyN, Universidad de Buenos Aires, Argentina, vcotik@dc.uba.ar
[10] Instituto de Investigación en Ciencias de la Computación (ICC), CONICET-UBA, Argentina
[11] Hospital de Pediatría 'Prof. Dr. Juan P. Garrahan', Argentina, dfilippo@gmail.com
[12] CONICET, Argentina
[13] German Research Center for Artificial Intelligence (DFKI), Germany, roland.roller@dfki.de
[14] Institut de Lingüística Aplicada, Universitat Pompeu Fabra, Spain, jorge.vivaldi@upf.edu

Abstract. In this paper, we provide an overview of the ninth annual edition of the CLEF eHealth evaluation lab. CLEF eHealth 2021 continues our evaluation resource building efforts around the easing and support of patients, their next-of-kins, health care professionals, and health scientists in understanding, accessing, and authoring electronic health information in a multilingual setting. The 2021 lab offered two tasks: Task 1 on multilingual Information Extraction (IE), this year extending

* With equal contribution, HS, LG & LK co-chaired the lab. Task 1 was led by VC and LAA, and organized by LAA, VC, DF, FL, RR, and JV; Task 2 was led by LG, GP, and HS, and organized by EB, NB-S, LG, GG-S, LK, PM, GP, HS, SS, RU, MV, and CX.

to a corpus of Spanish radiology reports; and Task 2 on Consumer Health Search (CHS) that builds on the previous year's Information Retrieval (IR) tasks. In total, 11 teams took part in these tasks (7 in Task 1 on IE and 4 in Task 2 on IR). Herein, we describe the resources created for these tasks and the evaluation methodology adopted, and we provide provide a brief summary of the participants of this year's challenges as well as the results obtained. As in previous years, the organizers have made data, tools, and more specific overview papers associated with the lab tasks available for future research and development.

Keywords: Entity Linking, Evaluation, Health Records, Information Extraction, Information Retrieval, Medical Informatics, Self-Diagnosis, Test-set Generation, Text Classification, Text Segmentation

1 Introduction

In recent years, electronic health (eHealth) content has become available 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 such medical content to support their clinical judgement and decision-making. The increasing difficulties experienced by these stakeholders in retrieving and digesting valid and relevant information in their preferred language to make health-centred decisions has motivated CLEF eHealth to organise yearly shared challenges since 2013.

More specifically, CLEF eHealth[15] was established as a lab workshop in 2012 as part of the Conference and Labs of the Evaluation Forum (CLEF, formerly known as Cross-Language Evaluation Forum). Since 2013 it has offered evaluation labs in the fields of layperson and professional health information extraction (IE), management, and retrieval (IR) with the aims of bringing together researchers working on related information access topics and providing them with datasets to work with and validate the outcomes. These labs and their subsequent workshops target:

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

[15] `https://clefehealth.imag.fr/`

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 narratives, 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 IE and IR in health care.

The ninth annual CLEF eHealth evaluation lab, CLEF eHealth 2021, aiming to build upon the resource development and evaluation approaches proposed in the previous years of the lab [40,19,10,18,11,39,20,15], offered the following two tasks:

- *Task 1.* Multilingual IE [3] and
- *Task 2.* Consumer Health Search (CHS) [16].

The *Multilingual IE* task builds upon the six previous editions of the task (2015–2020) which already addressed the analysis of biomedical text in English, French, Hungarian, Italian, Spanish, and German [29,27,28,30,31,25]. This year, the task focuses on Named Entity Recognition in Spanish ultrasound reports. Ten different classes of concepts in the radiology domain are distinguished, including Anatomical Entities, and Findings, that describe a pathological or abnormal event, negations, and indicators of probability or future outcomes. As well as complex entities, the task includes the challenge of semantic split of the dataset. That is, training, development, and test sets cover different semantic fields. This allows for a more realistic held-out evaluation.

The *Consumer Health Search* task is a continuation of the previous CLEF eHealth IR tasks that ran in 2013–2018, and 2020 [7,9,33,42,34,17,8,14]. It embraces 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 2021 task generates a new representative web corpus and collection of layperson medical queries. The task is structured into a number of optional subtasks as follows: (1) ad-hoc search, (2) weakly-supervised IR, and (3) document credibility assessment.

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

2 Materials and Methods

In this section, we describe the materials and methods used in the two tasks of the CLEF eHealth evaluation lab 2021. After specifying our text documents to process in Section 2.1, we address the human annotations, queries, and relevance assessments in Section 2.2. Finally, in Section 2.3, we introduce our evaluation methods.

2.1 Text Documents

Task 1. The dataset for this task consists of a corpus of Spanish radiology reports, more concretely pediatric ultrasounds from an Argentinian public hospital. These reports are generally written within a hospital information system by direct typing in a computer and are informed in only one section, where the most relevant findings are described. They are written using standard templates that guide physicians on the structure of the report when the findings are normal, but most of the time they are written in free text to be able to describe the findings discovered in abnormal studies. This fact results in great variations in both size and content of the reports, ranging from 8 to 193 words. Also, there are misspellings and inconsistencies in the usage of abbreviations, punctuation, and line breaks, as can be seen in Figure 1.

```
2a.
HIGADO de forma, tamano y ecoestructura normal.
VIA BILIAR intra y extrahepatica: no dilatada.
VESICULA BILIAR: de paredes finas sin imagenes endoluminales.
BAZO: tamano y ecoestructura normal.
Diametro longitudinal: 6.89 ( cm ) RETROPERITONEO VASCULAR: sin
alteraciones.
No se detectaron adenomegalias.
Ambos rinones de formsa, tamano y situacion habitual.
Adecuada diferenciacion cortico-medular.
RD Diam Long: 5.8 cm RI Diam long: 6.1 cm Vejiga de caracteristicas
normales.
No se observo liquido libre en cavidad abdomino-pelviana.
```

2y.
LIVER of regular form, size and echostructure.
Intra and extrahepatic BILE DUCT: non-dilated.
GALLBLADDER: thin walls and no endoluminal images.
SPLEEN: regular size and echostructure.
Longitudinal diameter: 6.89 (cm) VASCULAR RETROPERITONEAL: no alterations.
No adenomegalies were found.
Both kidneys of regular form, size and location.
Adequate corticomedullary differentiation.
RK Long diam: 5.8 cm LK Long diam: 6.1 cm Bladder of regular characteristics.
No free liquid was observed within the abdomino-pelvian cavity.

Fig. 1. A sample report, with its translation. It shows abbreviations ("RD" for right kidney, "RI" for left kidney, "Diam" for diameter), typos ("formsa" for "forma"), and inconsistencies (capitalization of "Vejiga" because of start of sentence without a full stop.)

Task 2. The document corpus used in the CHS task consists of web pages acquired from the CommonCrawl dump of 2021-04[16]. An initial list of websites was acquired from the 2018 CHS task which was built by submitting a set of medical queries to the Microsoft Bing Application Programming Interfaces (through the Azure Cognitive Services) repeatedly over a period of a few weeks, and acquiring the uniform resource Locators (URL) of the retrieved results. The domains of the acquired URLs were then included in the list, except some domains that were excluded for decency reasons. The list was augmented by including a number of known reliable and unreliable health websites, and social media contents of ranging reliability levels, from lists previously compiled by health institutions and agencies [17]. From this initial list of domains, a sample of domains was identified for final acquisition. This list was further extended by including websites, which were highly relevant for the task queries to create the final domain list with 600 domains. This introduced 13 new domains compared to the 2018 collection, and all domains were newly crawled from the latest CommonCrawl 2021-04.

The corpus was complemented with social media documents from *Reddit* and *Twitter*. A list of 150 health topics related to various health conditions was selected. Search queries were manually generated from those topics and were submitted to Reddit to retrieve posts and comments. The same process was applied on Twitter to get related tweets from the platform. A social media document was defined as a text obtained by a single interaction, therefore for Reddit one document is composed by a post, one comment of the post and associated meta-information. For Twitter, a document is a single tweet with its associated meta-information.

2.2 Human Annotations, Queries, and Relevance Assessments

Task 1. The radiology text data is annotated with seven different classes of entities: *Finding, Anatomical Entity, Location, Measure, Degree, Type of Measure* and *Abbreviation*. Additionally, hedges are also identified, distinguishing *Negation, Uncertainty and Conditional Temporal*. An example annotation can be found in Figure 2, and the frequency of each type of entity can be seen in Figure 3.

The phenomena under study have some challenging properties. For example, entities can be embedded within other entities. Moreover, entities can be discontinuous, and they can even span over sentence boundaries. The entity type *Finding* is particularly challenging, as it presents great variability in its textual forms. It ranges from a single word to more than ten words in some cases, comprising all kinds of phrases. However, this is also the most informative type of entity for the potential users of these annotations. Another challenging phenomena is the regular polysemy observed between *Anatomical entities* and *Locations*. In the manual annotation process, we have found that human annotators have

[16] `https://commoncrawl.org/2021/02/january-2021-crawl-archive-now-available/`

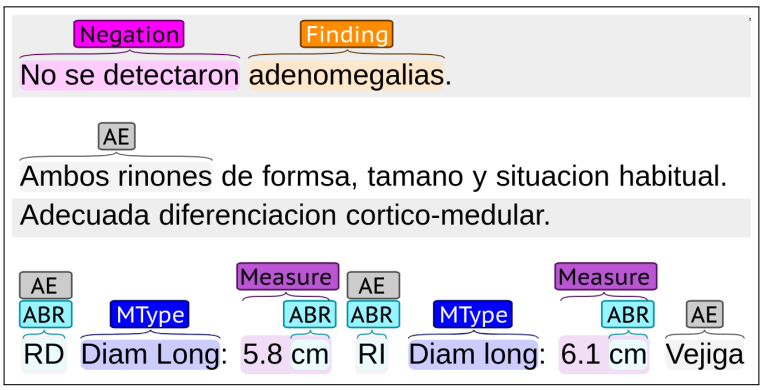

Fig. 2. A snippet of the report in Figure 1, with manual annotations. Abbreviations: AE — Anatomical Entity, ABR — Abbreviation, MType — Type of Measure.

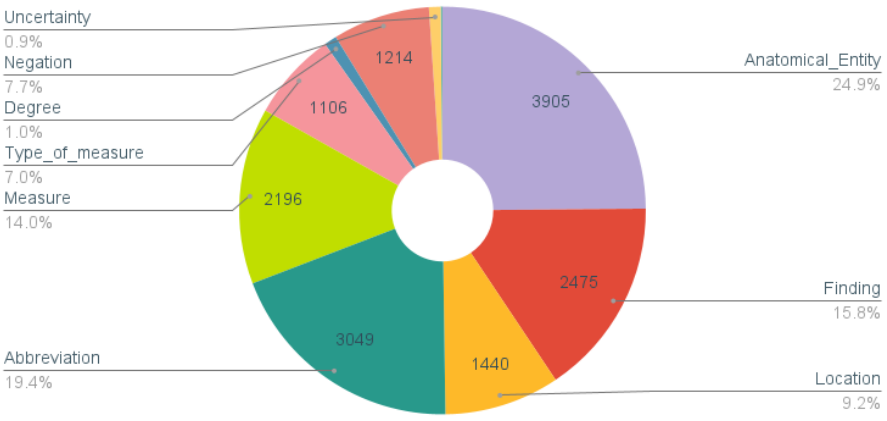

Fig. 3. Number and frequency of occurrences of the different kinds of entities in the annotated dataset for Task 1.

less agreement on those categories than on the rest, and automatic classifiers also experience difficulties to consistently classify those as well.

The given corpus consists of a total of 513 ultrasound reports, with 35,000 words and over 15,000 annotated named entities. In order to assess the portability of the approaches, half of the reports were provided as training, and the other half for testing, making sure that the testing partition contained portions of text that belonged to previously unseen phenomena. Reports were manually annotated by clinical experts [4] and then revised by linguists. Annotation guidelines and training were provided for both rounds of annotation. More information about the dataset can be found in [4]. Nevertheless, for the challenge the annotation criteria has been reviewed and some annotations have been modified.

The task, called SpRadIE (for Spanish Radiology Information Extraction), was inspired by previous research on this subject [5,2].

Task 2. The CHS task, Task 2, used a new set of 55 queries in English for realistic search scenarios. The queries were constructed either by hand, based on research interests and expertise of the organizers on multiple sclerosis and diabetes, or by using searches issued by the public to social media search services. Namely, the queries were manually authored and tailored by experts from established search scenarios and manually selected from a list of Google trends related queries to best fit each automatically extracted search scenario from social media (e.g., Twitter and Reddit).

Each query was manually labelled by the organizers with a narrative in English to describe the search intent or to capture the submission text for manually created queries and social medial queries, respectively. To illustrate, some queries and narratives appear as follows:

- Scenario 22:
  - Query: *my risk for developing type 2 diabetes*
  - Narrative: *You read that the risk for developing type 2 diabetes is increasing due to environmental and lifestyle factors, and you want to know more about your own risk.*
- Scenario 68:
  - Query: *List of multiple sclerosis symptoms*
  - Narrative: *I am a 40 year old patient with MS, and I have very vage symptoms, including fatigue, brain fog, foot drop, difficulties passing urine, problems turning right. Are these related to MS or might I have another disease in addition?*
- Scenario 105:
  - Query: *wisdom tooth cuts gum pain*
  - Narrative: *Hi all My wisdom tooth is currently cutting it's way through my bottom right gum the pain is intense throbbing aching jaw and weirdly a sore throat especially when swallowing. I just wonder if this is normal as I've had two wisdom tooth come through before with no pain at all. Thank you!*

People with lived experience of the related medical conditions were consulted to motivate, validate, and refine the narratives. Furthermore, the queries were enriched by the organizers to have a theme (manually created ones) or name (social media) to ease classifying them, but these were neither released to the participants nor used for evaluation.

The subtasks 1, 2, and 3 used these 55 queries with 5 released for training and 50 reserved for testing; the test topics contained a balanced sample of the manually constructed and automatically extracted search scenarios.

Relevance assessments are currently in progress and will be detailed in the CHS task overview [16]. Similar to the 2016 and 2017 pools, we created the pool using the rank-biased precision (RBP)-based Method A (Summing contributions) [26] 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 a study published in 2007 on RBP [35]). This strategy, named RBPA has been proven more efficient than traditional fixed-depth or stratified pooling to evaluate systems under fixed assessment budget constraints [22], as it is the case for this task. All participants' runs were considered on the document's pool, along with six baselines provided by the organizers. In order to guarantee the judgements of the documents of the participants' runs, half of the pool is composed by their documents and half from documents of the baselines' runs.

Along with relevance assessments, readability and credibility judgments were also collected for the assessment pool; these were used to evaluate systems across different dimensions of relevance (see [12] for further information about the three dimensions).

The relevance, readability, and credibility assessments were performed by 26 volunteers in May–June 2021. Of these assessors, 16 were from Australia, 1 from Finland, 3 from France, 2 from Ireland, and 4 from Italy. The numbers of female and male assessors were 19 and 7, respectively. All assessors were recruited, trained, and supervised by the organizers by using bespoke written materials from April to June 2021. The recruitment took place on social media and via email, using both organizers' existing contacts and snowballing.

Assessments were implemented online by the organizers' expanding and customising the Relevation! tool for relevance assessments [21] to capture our task dimensions, scales, and other preferences (Figures 4, 5, and 6). Each assessor was initially assigned 2 queries to be assessed, and in the end, every assessor completed 1 to 4 queries. Each query was associated with 250 documents to be assessed with respect to their relevance, readability, and credibility.

Ethical approval (2021/013) was obtained from the Human Research Ethics Committee of the Australian National University. Each study participant provided informed consent.

2.3 Evaluation Methods

Task 1. Participants could submit up to 4 runs. Lenient and exact match precision, recall, and F1 score were calculated. Submissions were evaluated with

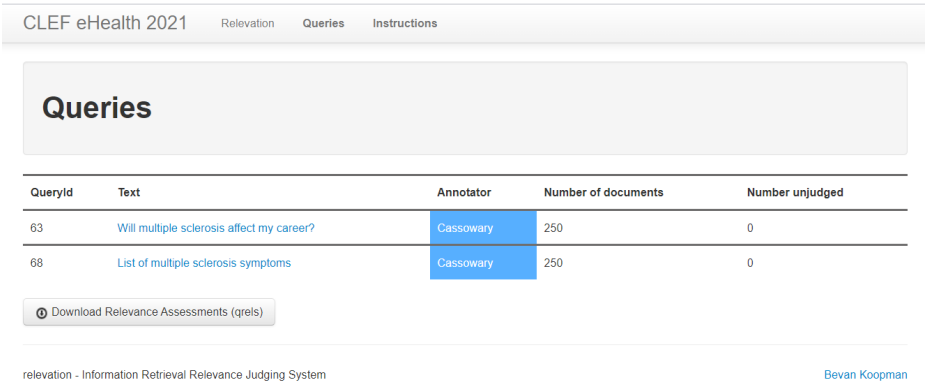

Fig. 4. CLEF eHealth Consumer Health Search Task 2021: Assessor's landing page

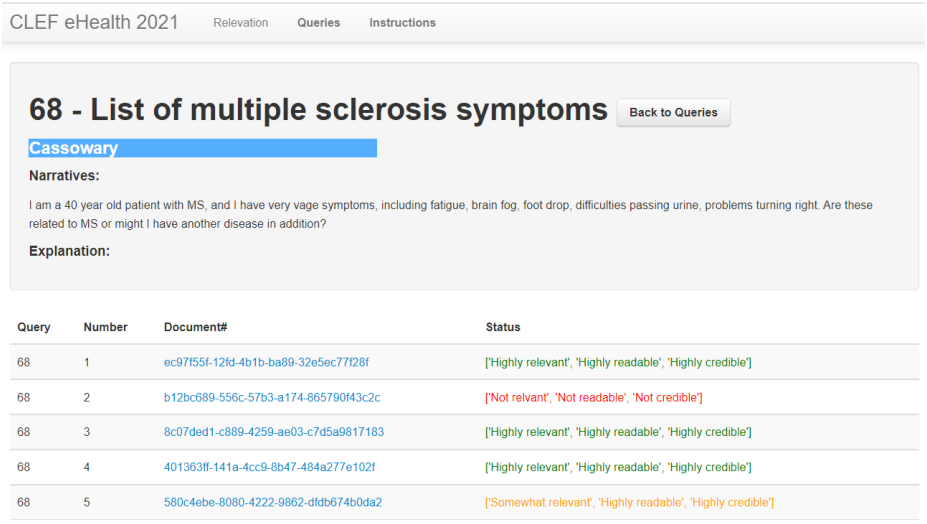

Fig. 5. CLEF eHealth Consumer Health Search Task 2021: Assessor's documents for a given query

micro-averaged lenient match F1. The lenient match is calculated using the Jaccard Index, as described in [13] and based on [1].

Task 2. For Subtasks 1, 2, and 3, participants could submit up to 4 runs in TREC format. Evaluation measures for Subtask 1, adhoc search task are Normalized Discounted Cumulative Gain (NDCG) at 10 (NDCG@10), BPref, and RBP, as well as other metrics adapted to other relevance dimensions such as uRBP and cRBP (with alpha value capturing the user expertise), an adapted metric to measure credibility relevance dimension based on uRBP. Subtask 3

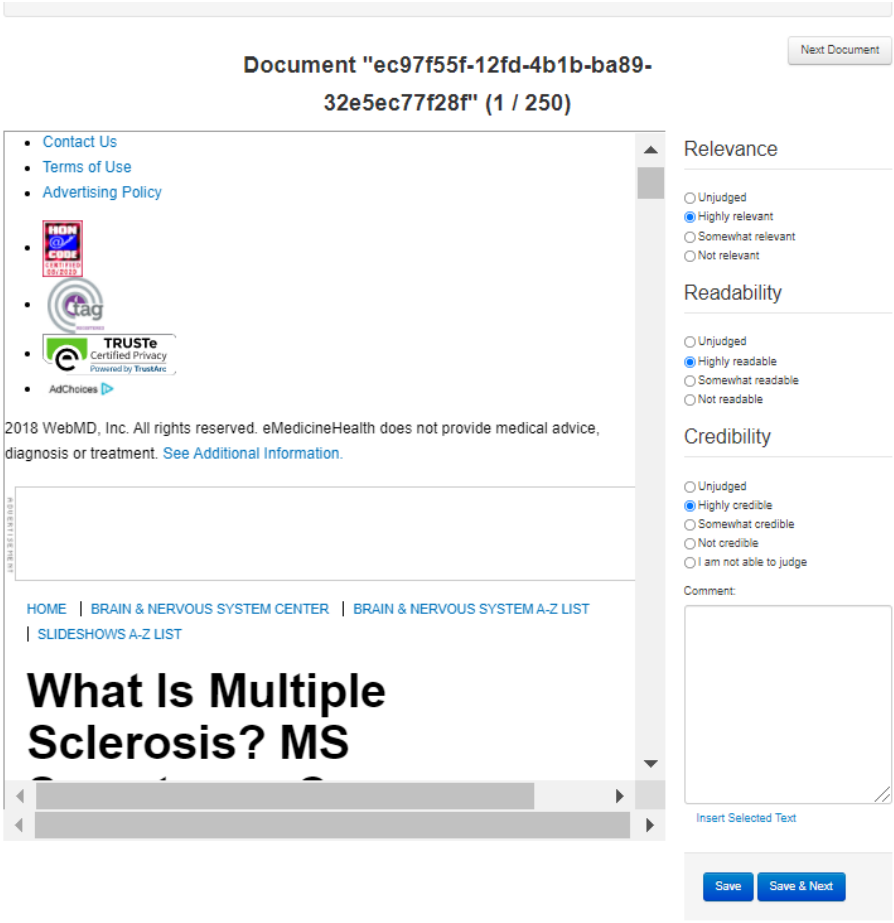

Fig. 6. CLEF eHealth Consumer Health Search Task 2021: Assessor's document view

used F1, Area under the receiver operating characteristic Curve (AUC), and Accuracy to measure a given system's ability to predict document credibility.

3 Results

The number of teams who registered their interest in CLEF eHealth 2021 Tasks 1 and 2 was 58 and 43 (and a total of 67 unique teams). In total, 7 and 4 teams submitted to the two shared tasks, respectively.

Task 1 Overall seven different teams participated in our shared task. Most prominent were participants from Spain, but also from Italy, UK and Colombia. Most participating teams were experimenting with different variations of neural

networks, particularly transformer-based approaches [37,23,41,36], but also bi-LSTMs [6]. Besides the challenge also includes submissions of a CRF [24], and a pattern based approach [32]. Overall, overlapping and discontinuous entities of the given dataset were the biggest challenge of the dataset, which made pre- and post-processing steps necessary. Moreover, in order to deal with the overlapping entities appropriately, the two highest scored teams make usage of multiple classifiers.

Table 1 shows the best result of each team's run. Best lenient precision, recall, and F1 are written in bold.

Table 1. Overall results for the best performing system for each team on the SpRadIE task, sorted by lenient micro-averaged F1.

	lenient			exact		
Team	PREC	REC	F1	PREC	REC	F1
EdIE (UnEd, UK) – run2	87.24	**83.85**	**85.51**	81.88	78.70	80.26
LSI (UNED, Spain) – run1	**90.28**	78.33	83.88	86.17	74.76	80.07
CTB (UPM, Spain) – run3	78.62	78.32	78.47	73.27	72.99	73.13
HULAT (UC3M, Spain) – run1	78.38	73.08	75.64	67.28	62.73	64.92
SINAI (UJaen, Spain) – run2	86.07	64.43	73.70	79.37	59.42	67.96
SWAP (UniBA, Italy) – run1	70.18	51.14	59.17	56.75	41.35	47.84
IMS (UniPD, Italy) – run1	9.29	57.62	16.00	5.45	33.77	9.38

The variation of the performance of the different systems across different kinds of entities can be seen in the boxplots in Figure 7. We can see that, although there is much variation in performance across systems (hence the long boxes), for some entities performance is lower, mostly those with fewer examples. Interestingly, types of entity with a big number of examples, like Location, still have low performance, for example, if compared with Anatomical Entities. It is interesting to see how performance for Abbreviations is very varied across approaches.

Task 2 had 4 teams submitting runs: In Subtask 2.1 on Ad Hoc IR, a 4-member team from the School of Computer Science, Zhongyuan University of Technology (ZUT) in Zhengzhou, China and a team with two members from the Information Management Systems (IMS) Research Group of the Italian University of Padova (UniPd) submitted runs. In Subtasks 2.2 on Weakly Supervised IR and 2.3 on Document Credibility Prediction, the leader of this IMS UniPd team, who has been a regular participant in previous CLEF eHealth IR tasks, submitted runs. Participants submissions were due by May 8th 2021 and 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 [16].

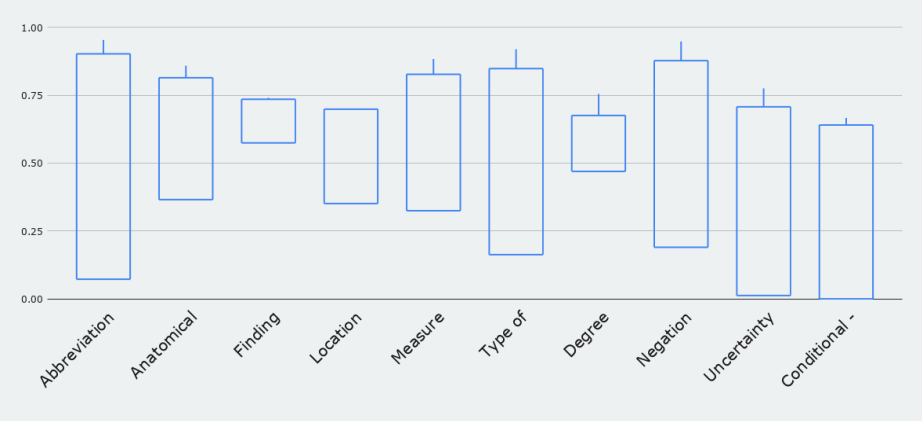

Fig. 7. Variation in the performance of different systems across different kinds of entities.

4 Conclusions

This paper provided an overview of the CLEF eHealth 2021 evaluation lab. The CLEF eHealth workshop series was established in 2012 as a scientific workshop with an aim of establishing an evaluation lab [38]. Since 2013, this annual workshop has been supplemented with two or more preceding shared tasks each year. In other words, they are the CLEF eHealth 2013–2020 evaluation labs [40,19,10,18,11,39,20,15]. These labs have offered a recurring contribution to the creation and dissemination of text analytics resources, methods, test collections, and evaluation benchmarks in order to ease and support patients, their next-of-kins, clinical staff, and health scientists in understanding, accessing, and authoring eHealth information in a multilingual setting.

In 2021, the CLEF eHealth lab offered two shared task. The first task was on multilingual IE and the second task was on CHS. These tasks built on the IE and IR tasks offered by the CLEF eHealth lab series since its inception in 2013. Test collections generated by these shared 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 the systems evaluated on the collections. These established CLEF IE and IR tasks 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 were used and all participants gained access to the test data at the same time, following which no further updates to systems were allowed. Shortly after releasing the test data (without labels or other solutions), the participating teams submitted their outputs from the frozen systems to the task organizers, who evaluated these results and reported the resulting benchmarks to the community.

The annual CLEF eHealth workhops and evaluation labs have matured and established their presence in 2012–2021. In total, 67 unique teams registered their interest and 11 teams took part in the 2021 tasks (7 in Task 1 on IE and 4 in Task 2 on IR). 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[17].

Acknowledgements

The CLEF eHealth 2021 evaluation lab has been supported in part by the CLEF Initiative. It has also been supported in part by the Our Health in Our Hands (OHIOH) initiative of the Australian National University (ANU), as well as the ANU School of Computing, ANU Research School of Population Health, and Data61/Commonwealth Scientific and Industrial Research Organisation. OHIOH is a strategic initiative of the ANU which aims to transform health care by developing new personalised health technologies and solutions in collaboration with patients, clinicians, and health care providers. Moreover, the lab has been supported in part by the bi-lateral Kodicare (Knowledge Delta based improvement and continuous evaluation of retrieval engines) project funded by the French ANR (ANR-19-CE23-0029) and Austrian FWF. We are also thankful to the people involved in the annotation, query creation, and relevance assessment exercises. 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. Bossy, R., Golik, W., Ratkovic, Z., Bessières, P., Nédellec, C.: BioNLP shared task 2013 — an overview of the Bacteria Biotope Task. In: Proceedings of the BioNLP Shared Task 2013 Workshop. pp. 161–169. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013)
2. Cotik, V.: Information extraction from Spanish radiology reports. In: PhD Thesis (2018)
3. Cotik, V., Alemany, L.A., Filippo, D., Luque, F., Roller, R., Vivaldi, J., Ayach, A., Carranza, F., Francesca, L.D., Dellanzo, A., Urquiza, M.F.: Overview of CLEF eHealth Task 1 - SpRadIE: A challenge on information extraction from Spanish Radiology Reports. In: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (September 2021)
4. Cotik, V., Filippo, D., Roller, R., Uszkoreit, H., Xu, F.: Annotation of Entities and Relations in Spanish Radiology Reports. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 177–184 (2017)
5. Cotik, V., Rodríguez, H., Vivaldi, J.: Spanish named entity recognition in the biomedical domain. In: Annual International Symposium on Information Management and Big Data. pp. 233–248. Springer (2018)

[17] `http://clef-ehealth.org/`

6. Fabregat, H., Duque, A., Araujo, L., Martinez-Romo, J.: LSI_UNED at CLEF eHealth2021: Exploring the effects of transfer learning in negation detection and entity recognition in clinical texts. In: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2021)
7. Goeuriot, L., Jones, G.J., Kelly, L., Leveling, J., Hanbury, A., Müller, H., Salanterä, S., Suominen, H., Zuccon, G.: ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. CLEF 2013 Online Working Notes 8138 (2013)
8. Goeuriot, L., Jones, G.J., Kelly, L., Leveling, J., Lupu, M., Palotti, J., Zuccon, G.: An Analysis of Evaluation Campaigns in ad-hoc Medical Information Retrieval: CLEF eHealth 2013 and 2014. Springer Information Retrieval Journal (2018)
9. Goeuriot, L., Kelly, L., Lee, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Gareth J.F. Jones, H.M.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes. Sheffield, UK (2014)
10. Goeuriot, L., Kelly, L., Suominen, H., Hanlen, L., Névéol, A., Grouin, C., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2015. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. Springer Berlin Heidelberg (2015)
11. Goeuriot, L., Kelly, L., Suominen, H., Névéol, A., Robert, A., Kanoulas, E., Spijker, R., Palotti, J., Zuccon, G.: CLEF 2017 eHealth Evaluation Lab overview. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 291–303. Springer Berlin Heidelberg (2017)
12. Goeuriot, L., Liu, Z., Pasi, G., Saez, G.G., Viviani, M., Xu, C.: 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)
13. Goeuriot, L., Suominen, H., Kelly, L., Alemany, L.A., Brew-Sam, N., Cotik, V., Filippo, D., Saez, G.G., Luque, F., Mulhem, P., Pasi, G., Roller, R., Seneviratne, S., Vivaldi, J., Viviani, M., Xu, C.: CLEF eHealth 2021 Evaluation Lab. In: Advances in Information Retrieval — 43st European Conference on IR Research. Springer, Heidelberg, Germany (2021)
14. Goeuriot, L., Suominen, H., Kelly, L., Liu, Z., Pasi, G., Gonzales, G.S., Viviani, M., Xu, C.: 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)
15. Goeuriot, L., Suominen, H., Kelly, L., Miranda-Escalada, A., Krallinger, M., Liu, Z., Pasi, G., Gonzalez Saez, G., Viviani, M., Xu, C.: Overview of the CLEF eHealth evaluation lab 2020. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 255–271. Springer International Publishing, Cham (2020)
16. Goeuriot, L., Suominen, H., Pasi, G., Bassani, E., Brew-Sam, N., González-Sáez, G., Kelly, L., Mulhem, P., Seneviratne, S., Gyanendra Upadhyay, R., Viviani, M., Xu, C.: Consumer Health Search at CLEF eHealth 2021. In: CLEF 2021 Evaluation Labs and Workshop: Online Working Notes. CEUR-WS (September 2021)
17. Jimmy, Zuccon, G., Palotti, J.: 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)

18. Kelly, L., Goeuriot, L., Suominen, H., Névéol, A., Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2016. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 255–266. Springer Berlin Heidelberg (2016)
19. Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2014. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 172–191. Springer Berlin Heidelberg (2014)
20. Kelly, L., Suominen, H., Goeuriot, L., Neves, M., Kanoulas, E., Li, D., Azzopardi, L., Spijker, R., Zuccon, G., Scells, H., Palotti, J.: Overview of the CLEF eHealth Evaluation Lab 2019. In: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Heinatz Bürki, G., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 322–339. Springer International Publishing, Cham (2019)
21. Koopman, B., Zuccon, G.: Relevation!: an open source system for information retrieval relevance assessment. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. pp. 1243–1244. ACM (2014)
22. Lipani, A., Palotti, J., Lupu, M., Piroi, F., Zuccon, G., Hanbury, A.: Fixed-cost pooling strategies based on IR evaluation measures. In: European Conference on Information Retrieval. pp. 357–368. Springer (2017)
23. López-Úbeda, P., Díaz-Galiano, M.C., Ureña-López, L.A., Martín-Valdivia, M.T.: Pre-trained language models to extract information from radiological reports. In: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2021)
24. Ángel Martín-Caro García-Largo, M., Bedmar, I.S.: Extracting information from radiology reports by Natural Language Processing and Deep Learning. In: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2021)
25. 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)
26. Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. 27(1), 2:1–2:27 (Dec 2008)
27. Névéol, A., Cohen, K., Grouin, C., Hamon, T., Lavergne, T., Kelly, L., Goeuriot, L., Rey, G., Robert, A., Tannier, X., Zweigenbaum, P.: Clinical information extraction at the CLEF eHealth evaluation lab 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Working Notes. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-1609/ (2016)
28. Névéol, A., Anderson, R.N., Cohen, K.B., Grouin, C., Lavergne, T., Rey, G., Robert, A., Zweigenbaum, P.: CLEF eHealth 2017 multilingual information extraction task overview: ICD10 coding of death certificates in English and French. In: CLEF 2017 Online Working Notes. CEUR-WS (2017)
29. Névéol, A., Grouin, C., Tannier, X., Hamon, T., Kelly, L., Goeuriot, L., Zweigenbaum, P.: CLEF eHealth evaluation lab 2015 task 1b: clinical named entity recognition. In: CLEF 2015 Online Working Notes. CEUR-WS (2015)

30. Névéol, A., Robert, A., Grippo, F., Morgand, C., Orsi, C., Pelikán, L., Ramadier, L., Rey, G., Zweigenbaum, P.: CLEF eHealth 2018 multilingual information extraction task overview: ICD10 coding of death certificates in French, Hungarian and Italian. In: CLEF 2018 Online Working Notes. CEUR-WS (2018)
31. Neves, M., Butzke, D., Dörendahl, A., Leich, N., Hummel, B., Schönfelder, G., Grune, B.: 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)
32. Nunzio, G.M.D.: IMS-UNIPD @ CLEF eHealth Task 1: A Memory Based Reproducible Baseline. In: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2021)
33. Palotti, J., Zuccon, G., Goeuriot, L., Kelly, L., Hanburyn, A., Jones, G.J., Lupu, M., Pecina, P.: CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information about Medical Symptoms. In: CLEF 2015 Online Working Notes. CEUR-WS (2015)
34. Palotti, J., Zuccon, G., Jimmy, Pecina, P., Lupu, M., Goeuriot, L., Kelly, L., Hanbury, A.: CLEF 2017 Task Overview: The IR Task at the eHealth Evaluation Lab. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2017)
35. 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)
36. Polignano, M., de Gemmis, M., Semeraro, G.: Comparing Transformer-based NER approaches for analysing textual medical diagnoses. In: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2021)
37. Solarte-Pabón, O., Montenegro, O., Blazquez-Herranz, A., Saputro, H., Rodriguez-González, A., Menasalvas, E.: Information Extraction from Spanish Radiology Reports using multilingual BERT. In: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2021)
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. vol. 1178. CEUR Workshop Proceedings (CEUR-WS.org) (2012)
39. Suominen, H., Kelly, L., Goeuriot, L., Névéol, A., Ramadier, L., Robert, A., Kanoulas, E., Spijker, R., Azzopardi, L., Li, D., Jimmy, Palotti, J., Zuccon, G.: Overview of the CLEF eHealth Evaluation Lab 2018. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 286–301. Springer Berlin Heidelberg (2018)
40. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Pradhan, S., South, B.R., Mowery, D.L., Jones, G.J., Leveling, J., Kelly, L., Goeuriot, L., Martinez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization, pp. 212–231. Springer Berlin Heidelberg (2013)
41. Suárez-Paniagua, V., Dong, H., Casey, A.: A multi-BERT hybrid system for Named Entity Recognition in Spanish radiology reports. In: CLEF eHealth 2021. CLEF 2021 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (2021)
42. Zuccon, G., Palotti, J., Goeuriot, L., Kelly, L., Lupu, M., Pecina, P., Mueller, H., Budaher, J., Deacon, A.: The IR Task at the CLEF eHealth Evaluation Lab 2016: User-centred Health Information Retrieval. In: CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS (September 2016)