Overview of the CLEF eHealth Evaluation Lab 2015

Lorraine Goeuriot[1], Liadh Kelly[2], Hanna Suominen[3], Leif Hanlen[4]
Aurélie Névéol[5], Cyril Grouin[5], João Palotti[6], and Guido Zuccon[7]⋆

[1] LIG, Université Grenoble Alpes, France, Lorraine.Goeuriot@imag.fr
[2] ADAPT Centre, Trinity College, Dublin, Ireland Liadh.Kelly@tcd.ie
[3] NICTA, The Australian National University, University of Canberra, and University of Turku, Canberra, ACT, Australia, firstname.lastname@nicta.com.au
[4] NICTA, The Australian National University, and University of Canberra, Canberra, ACT, Australia, firstname.lastname@nicta.com.au
[5] LIMSI CNRS UPR 3251, France, firstname.lastname@limsi.fr
[6] Vienna University of Technology, Austria palotti@ifs.tuwien.ac.at
[7] Queensland University of Technology, Australia g.zuccon@qut.edu.au

Abstract. This paper reports on the 3rd CLEFeHealth evaluation lab, which continues our evaluation resource building activities for the medical domain. In this edition of the lab, we focus on easing patients and nurses in authoring, understanding, and accessing eHealth information.
The 2015 CLEFeHealth evaluation lab was structured into two tasks, focusing on evaluating methods for information extraction (IE) and information retrieval (IR). The IE task introduced two new challenges. Task 1a focused on clinical speech recognition of nursing handover notes; Task 1b focused on clinical named entity recognition in languages other than English, specifically French. Task 2 focused on the retrieval of health information to answer queries issued by general consumers seeking information to understand their health symptoms or conditions.
The number of teams registering their interest was 47 in Tasks 1 (2 teams in Task 1a and 7 teams in Task 1b) and 53 in Task 2 (12 teams) for a total of 20 unique teams. The best system recognized 4, 984 out of 6, 818 test words correctly and generated 2, 626 incorrect words (i.e., 38.5% error) in Task 1a; had the F-measure of 0.756 for plain entity recognition, 0.711 for normalized entity recognition, and 0.872 for entity normalization in Task 1b; and resulted in P@10 of 0.5394 and nDCG@10 of 0.5086 in Task 2. These results demonstrate the substantial community interest and capabilities of these systems in addressing challenges faced by patients and nurses. As in previous years, the organizers have made data and tools available for future research and development.

Keywords: Evaluation, Information Retrieval, Information Extraction, Medical Informatics, Nursing Records, Patient Handoff/Handover, Speech Recognition, Test-set Generation, Text Classification, Text Segmentation, Self-Diagnosis

⋆ In alphabetical order, LG & LK co-chaired the lab. In order of contribution, HS & LH led Task 1a. In order of contribution, AN & CG led Task 1B. In alphabetical order, JP & GZ led Task 2.

1 Introduction

This paper presents an overview of the CLEFeHealth 2015 evaluation lab[8], organized within the Conference and Labs of the Evaluation Forum (CLEF)[9] to support development of approaches, which support patients, their next-of-kins, and clinical staff in understanding, accessing and authoring health information. This third year of the evaluation lab aimed to build upon the resource development and evaluation approaches offered in the first two years of the lab, which focused on patients and their next-of-kins' ease in understanding and accessing health information.

The first CLEFeHealth lab [1] contained three tasks: Task 1 on named entity recognition and/or normalization of disorders [2]; Task 2 on acronyms/ abbreviations [3] in clinical reports; Task 3 health-focused web information retrieval, supporting laypeople's information needs stemming from clinical reports [4].

The second CLEFeHealth [5] expanded our year-one efforts and again organized three tasks. Specifically, Task 1 aimed to help patients (or their next-of-kin) by addressing visualisation and readability issues related to their hospital discharge documents and related information search on the Internet [6]. Task 2 continued the IE work of the 2013 CLEFeHealth lab, specifically focusing on IE of disorder attributes from clinical text [7]. Task 3 further extended the 2013 IR task, with a cleaned version of the 2013 document collection being produced and the introduction of a new query generation method, as well as multilingual queries [8].

The 2015 lab was split into two tasks focusing on information extraction and information retrieval. The IE task introduced two new challenges: Task 1a focused on clinical speech recognition (SR) of nursing shift changes [9]; Task 1b focused on named entity recognition in clinical reports in languages other than English, specifically French clinical reports [10]. The IR task focused on a new type of queries people issue to obtain information on the web [11]; Task 2a considered English queries, while Task 2b considered multilingual queries obtained through expert translation of the English queries[10].

In total the 2015 edition of the CLEFeHealth lab attracted 20 teams to submit 4 submissions[11] to Task 1a, 38 to Task 1b, and 97 to Task 2; demonstrated the capabilities of these systems in contributing to patients and nurses' understanding and information needs; and made data, guidelines, and tools available for future research and development. The lab workshop was held at CLEF in September 2015.

[8] `https://sites.google.com/site/clefehealth2015/`

[9] `http://www.clef-initiative.eu/`

[10] In the remaining we will refer to Task 2a as Task 2; we will use Task 2b to refer to the multilingual queries only when this specific case was considered. Note that only one team submitted runs for multilingual queries.

[11] Note that in this paper, we refer to submissions, systems, experiments, and runs as *submissions*.

2 Tasks Motivations

2.1 Task 1

Laypeople find health related documents to be difficult to understand; clinicians have also problems in understanding the jargon of other professional groups even though policies and regulations emphasise the need to document care in a comprehensive manner and provide further information on health conditions to help their understanding. An example from a US discharge document is "*AP: 72 yo f w/ ESRD on HD, CAD, HTN, asthma p/w significant hyperkalemia & associated arrythmias*". Another example from a French hospital stay report is "*FOGD sous A.G. + dilatation chez un patient porteur d'un carcinome épidermoide du 1/3 supérieur de l'oesophage T2N0M0 opéré en 97*". However, authors of both care documents and consumer leaflets are overloaded with information and face many challenges in the timely and efficient generation, processing and sharing of such information. One example here is clinical handover between nurses, where verbal handover and note taking can lead to loss of information. As described in [1], there is much need for techniques, which support individuals in understanding such clinical documents including in languages other than English. This edition of the CLEF eHealth lab answers the call for biomedical shared tasks in languages other than English [12] by introducing a task addressing clinical named entity recognition and normalization in biomedical documents in French.

In addition, auto-converting a verbal nursing handover to text and then highlighting important information within the transcription — or even filling out a structured handover form — for the next nurse would aid care documentation and release nurses time to, for example, discuss these resources and provide further information for a longer time with the patients. Task 1a aims at tackling this challenge.

2.2 Task 2

The use of the Web as source of health-related information is a wide-spread phenomena. Search engines are commonly used as a means to access health information available online. The 2013 and 2014 CLEFeHealth lab Task 3 aimed at evaluating the effectiveness of search engines to support people when searching for information about known conditions, e.g. to answer queries like "thrombocytopenia treatment corticosteroids length" [8, 4, 13]. Other types of searches for health related information are for self-diagnosis purposes, often issued before attending a medical professional (or to help the decision of attending) [14]. Previous research has shown that exposing people with no or scarce medical knowledge to complex medical language may lead to erroneous self-diagnosis and self-treatment and that access to medical information on the Web can lead to the escalation of concerns about common symptoms (e.g., cyberchondria) [15, 16]. Research has also shown that current commercial search engines are yet far from being effective in answering such queries [17]. We thus decided to investigate this type of queries in the 2015 CLEFeHealth lab Task 2. We expected these queries to pose a new challenge to the participating teams; a challenge that, if solved, would lead to significant contributions towards improving how current commercial search engines answer health queries.

3 Materials and Methods

3.1 Speech and Text Documents

The NICTA Synthetic Nursing Handover Data was used in Task 1a [18, 9]. This set of 200 synthetic patient cases (i.e., 100 for training and another 100 for testing) was developed for SR and IE related to nursing shift-change handover in 2012–2015. Each case consisted of a patient profile; a written, free-form text paragraph (i.e., the written handover document) to be used as a reference standard in SR; and its spoken (i.e., the verbal handover document) and speech-recognized counterparts.

For Task 1b, two types of biomedical documents were used: a total of 1,668 titles of scientific articles indexed in The MEDLINE database, and 6 full text drug monographs published by the European Medicines Agency (EMEA).

For Task 2, the CLEFeHealth 2014 Task 3 large crawl of health resources on the Internet was used. It contained about one million documents [19] and originated from the Khresmoi project[12]. The crawled domains were predominantly health and medicine sites, which were certified by the HON Foundation as adhering to the HONcode principles (appr. 60–70 per cent of the collection), as well as other commonly used health and medicine sites such as Drugbank, Diagnosia and Trip Answers.[13] Documents consisted of pages on a broad range of health topics and were targeted at both the general public and healthcare professionals. They were made available for download on the Internet in their raw HTML format along with their URLs to registered participants on a secure password-protected server.

3.2 Human Annotations, Queries, and Relevance Assessments

For Task 1b, the annotations covered ten types of entities of clinical interest, defined by Semantic Groups in the Unified Medical Language System (UMLS) [20]: *Anatomy*, *Chemicals & Drugs*, *Devices*, *Disorders*, *Geographic Areas*, *Living Beings*, *Objects*, *Phenomena*, *Physiology*, *Procedures*. The annotations marked each relevant entity mention in the documents, and assigned the corresponding semantic type(s) and Concept Unique Identifier(s) or CUIs. Each document was annotated by one professional annotator (two annotators participated in total) according to detailed guidelines [21]. The annotations were then validated and revised by a senior annotator to ensure annotation consistency and correctness

[12] Medical Information Analysis and Retrieval, `http://www.khresmoi.eu`

[13] Health on the Net, `http://www.healthonnet.org`, `http://www.hon.ch/HONcode/Patients-Conduct.html`, `http://www.drugbank.ca`, `http://www.diagnosia.com`, and `http://www.tripanswers.org`

throughout the corpus. The corpus was split evenly between training data supplied to the participants at the beginning of the lab, and an unseen test set used to evaluate participants' systems.

For Task 2, queries were obtained by showing images and videos related to medical symptoms to users, who were then asked which queries they would issue to a web search engine if they or their next-of-kins were exhibiting such symptoms and thus wanted to find more information to understand these symptoms or which condition they were affected by. This methodology for eliciting circumlocutory, self-diagnosis queries was shown to be effective by Stanton et al. [22]; Zuccon et al. [17] showed that current commercial search engines are yet far from being effective in answering such queries.

Following the methodology in [22, 17], 23 symptoms or conditions that manifest with visual or audible signs (e.g. ringworm or croup) were selected to be presented to users to collect queries. A cohort of 12 volunteer university students and researchers based in the organisers' institutions was used to generate the queries. A total of 266 possible unique queries were collected; of these, 67 queries (22 conditions with 3 queries and 1 condition with 1 query) were selected to be used in this year's task. In addition, we developed translations of this query set into Arabic (AR), Czech (CS), German (DE), Farsi (FA), French (FR), Italian (IT) and Portuguese (PT); these formed the multilingual query sets which were made available to participants for submission of multilingual runs. Queries were translated by medical experts available at the organisers institutions.

Relevance assessments were collected by pooling participants' submitted runs as well as baseline runs. Assessment was performed by four paid medical students who had access to the query the document was retrieved for, as well as the target symptom or condition that was used to obtained the query during the query generation phase. Along with relevance assessments, readability judgements were also collected for the assessment pool. Assessments were provided on a four point scale: 0, It is very technical and difficult to read and understand; 1, It is somewhat technical and difficult to read and understand; 2, It is somewhat easy to read and understand; 3, It is very easy to read and understand.

3.3 Evaluation Methods

In Task 1a, the participants needed to submit their processing results. Submissions that developed the SR engine itself were evaluated separately from those that studied post-processing methods for the speech-recognized text. Also a separate submission category was assigned to solutions based on both SR and text post-processing. Each participant was allowed to submit up to two systems to the first category and up to two systems to the second category. If addressing both these categories, the participant was asked to submit all possible combinations of these systems as their third category submission. Final submission then consisted of the processing outputs for each method on the 100 training and 100 test documents.

In Task 1b, teams could submit up to two runs for three subtasks that were evaluated separately on the two types of text supplied (MEDLINE and EMEA):

1/for **plain entity recognition**, raw text was supplied to participants who had to submit entity annotations comprising entity offsets and entity types. 2/for **normalized entity recognition**, raw text was supplied to participants who had to submit entity annotations comprising entity offsets, entity types, and entity normalization (UMLS CUIs). 3/for **entity normalization**, raw text and plain entity annotations were supplied to participants who had to submit entity normalization (UMLS CUIs). For each of the subtasks, the system output on the unseen test set was compared to the gold standard annotations and precision recall and F-measure was computed.

In Task 2, teams could submit up to ten runs for the English queries, and an additional ten runs for each of the multilingual query languages. Teams were required to number runs such as that run 1 was a baseline run for the team; other runs were numbered from 2 to 10, with lower numbers indicating higher priority for selection of documents to contribute to the assessment pool (i.e. run 2 was considered of higher priority than run 3).

Teams received data from November 2014 to April 2015. In Task 1a, teams could access the training documents on 15 November 2014 and test documents on 23 April 2015. In Tasks 1b, data was divided into training and test sets; the evaluation for these tasks was conducted using the blind, withheld test data (documents for Task 1b). Teams were asked to stop development as soon as they downloaded the test data. The training set and test set for Tasks 1b and the 5 example queries and the test queries for Task 2 were released from December 2014 and April 2015 respectively. For Task 1b, the test set was released in two steps because the plain entity gold standard was needed as an input for the normalization subtask. Participants had to submit their runs for the entity recognition subtasks before the entity gold standard was released. Evaluation results were announced to the participants for the three tasks in May.

In Task 2, for each query, the top 10 documents returned in runs 1, 2 and 3 produced by the participants[14] were pooled to form the relevance assessment pool. In addition, the organisers also generated baseline runs using BM25, TF-IDF and Dirichlet Language model, as well as a set of benchmark systems that ranked documents by estimating both (topical) relevance and readability[15]; these were pooled with the same methodology used for participants runs. A total of 8,713 documents were assessed.

The system performance in the different tasks was evaluated against task-specific criteria. In Task 1a, we challenged the participants to minimize the number of incorrectly recognized words on the independent test set. This correctness was evaluated on the entire test set using the primary measure of the

[14] With the exclusion of multilingual submissions, for which runs were not pooled due to the larger assessment effort pooling these runs would have required. Note that only one team submitted multilingual runs.

[15] Run 1: linear interpolation of BM25 scores (weight 0.9) and Dale Chall readability score (weight 0.1); run 2: multiplication of BM25 scores and log of word frequency extracted from Wikipedia; run 3: TF-IDF and Flesh-Kincaid readability scores combined via an inverse logarithmic function. See [11] for details.

percentage of incorrect words (aka the error rate percentage E) as defined by the Speech Recognition Scoring Toolkit (SCTK), 2.4.0 without punctuation as a differentiating feature. This measure sums up the percentages of substituted (S), deleted (D), and inserted (I) words (i.e., $E = S + D + I$). As secondary measures, we reported the percentage of correctly detected words (C) on the entire test set together with the breakdown of E to D, I, and S. We also documented the raw word numbers behind these percentages, provided more details on performance differences across the individual handover documents, and assessed the resubstitution performance on the training set. We used two baseline systems in Task 1a, namely Dragon Medical 11.0 and Majority, which assumed that the right number of words is detected and recognized every word as the most common training word with the correct capitalization. Statistical differences between the error rate percentages of the two baselines and participant submissions were evaluated using the Wilcoxon signed-rank test (W) [23]. After ranking the baselines and submissions based on their error rate percentage on the entire dataset for testing, W was computed for the paired comparisons from the best and second-best system to the second-worst and worst system. The resulting p value and the significance level of 0.05 was used to determine if the median performance of the higher-ranked method was significantly better than this value for the lower-ranked method. All statistical tests were computed using R 3.2.0.

Tasks 1b system performance was evaluated using precision, recall and F-measure. The official primary measure was exact match F-measure.

In Task 2, system evaluation was conducted using precision at 10 (p@10) and normalised discounted cumulative gain [24] at 10 (nDCG@10) as the primary and secondary measures, respectively. Precision was computed using the binary relevance assessments; nDCG was computed using the graded relevance assessments. A separate evaluation was conducted using both relevance assessments and readability assessments following the methods in [25]. For all runs, Rank biased precision (RBP)[16] was computed along with readability-biased modifications of RBP, namely uRBP (using the binary readability assessments) and uRBPgr (using the graded readability assessments). More details on the readability-based evaluation are provided in the Task overview paper [11].

The organizers provided the following evaluation tools on the Internet. To supplement the usage guidelines of SCTK, we provided the Task 1a participants with some helpful tips. More specifically, we released an example script for removing punctuation and formatting text files; a formatted reference file and Dragon baseline for the training set; overall and document-specific evaluation results for this file pair; and commands to perform these evaluations and ensure the correct installation of SCTK. For Task 1b, results were computed using the `brateval` [26] program which we extended to cover the evaluation of normalized entities. The updated version of `brateval` was supplied to task participants along with the training data. For Task 2, precision and nDCG were computed

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

using `trec_eval`; while the readability-biased evaluation was performed using `ubire`[17].

4 Results

The number of people who registered their interest in Tasks 1 and 2 was 47 and 53, respectively, and in total 20 teams with unique affiliations submitted to the shared tasks (Tables 1 and 2). No team participated in all tasks. Two teams participated in Tasks 1b and 2 (Table 2). Teams represented Argentina, Australia, Belarus, Botswana, Canada, China, Czech Republic, France, Germany, India, Korea, Spain, The Netherlands, Thailand, Tunisia, and Vietnam.

Table 1. Participating teams

ID	Team	Affiliation	Location
1	CISMeF	CISMeF, LITIS	France
2	CUNI	Institute of Formal and Applied Linguistics	Czech Republic
3	ECNU-ICA	Shanghai Key Laboratory of Multidimensional Information Processing	China
4	Erasmus	Erasmus Mc	Netherlands
5	FDUSGinfo	Fudan University	China
6	GRIUM	RALI, DIRO, University of Montreal	Canada
7	HCMUS	Vietnam National University	Vietnam
8	HIT-W	Harbin Institute of Technology	China
9	IHS-RD	IHS Inc	Belarus
10	KISTI	KISTI	Korea
11	KU-CS	Kasetsart University	Thailand
12	LIMSI-ILES	LIMSI	France
13	Miracl	Miracl Lab, IRIT	Tunisia, France
14	TUC-MI/MC	Technische Universität Chemnitz	Germany
15	UBML	University of Botswana	Botswana
16	UC	University of Canberra	Australia
17	UPF	Universitat Pompeu Fabra, Universidad de Buenos Aires	Spain, Argentina
18	USST	University of Shanghai for science and technology	China
19	Watchdogs	Dhirubhai Ambani Institute of Information and Communication Technology	India
20	YorkU	York University	Canada

In total 209 systems were submitted to the challenge (Table 2).

Task 1a opened in both verbal and written formats the total of 200 synthetic clinical documents that can be used for studies on nursing documentation and informatics. It attracted 48 team registrations with 21 teams confirming

[17] `https://github.com/ielab/ubire`, [25].

Table 2. The tasks that the teams participated in

ID	Team	Number of submitted systems per task				
		1a	1b	2a	2b	
1	CISMeF		4			
2	CUNI			10	70 (10 runs per language)	
3	ECNU-ICA			10		
4	Erasmus		12			
5	FDUSGinfo			10		
6	GRIUM			7		
7	HCMUS			8		
8	HIT-W		6			
9	IHS-RD		8			
10	KISTI			8		
11	KU-CS			4		
12	LIMSI-ILES		2	5		
13	Miracl			5		
14	TUC-MI/MC	4				
15	UBML			10		
16	UC	Rejected				
17	UPF		2			
18	USST			10		
19	Watchdogs		4			
20	YorkU			10		
	Systems:	4	38	97	70	*Total: 209*
	Teams:	1	7	12	1	

their participation through email. Two interdisciplinary teams submitted two SR methods each. Unfortunately, UC.2 submission was incomplete and thus was rejected by the organizers.

The Dragon baseline had clearly the best performance (i.e., $E = 38.5$) on the Task 1a test documents, followed by the TUC_MI/MC.2 ($E = 52.8$), TUC_MI/MC.1 ($E = 52.3$), UC.1 ($E = 93.1$), and the Majority baseline ($E = 95.4$). The performance of the Dragon baseline on the test set was significantly better than that of the second-best system (i.e., TUC_MI/MC.2, $W = 302.5$, $p < 10^{-12}$). However, this rank-2 system was not significantly better than the third-best method (i.e., TUC_MI/MC.1), but this rank-3 system was significantly better than the fourth-best system (i.e., UC.1, $W = 0$, $p < 10^{-15}$). Finally, the performance of the lowest-ranked system (i.e., the Majority baseline) was significantly worse than that of this rank-4 system ($W = 1,791.5$, $p < 0.05$). See the Task 1a [9] for more detailed evaluation results.

In total, seven teams submitted systems for Task 1b. For the plain entity recognition subtask, seven teams submitted a total of 10 runs for each corpus (EMEA and MEDLINE). For the normalized entity recognition task, four teams submitted a total of 5 runs for each corpus. For the normalization task, three teams submitted a total of 4 runs for each corpus. The best system had an F-measure of 0.756 for plain entity recognition, 0.711 for normalized entity recognition and 0.872 for entity normalization. See Tables 3, 4, 5, 6, 7 and 8 for details.

Table 3. Task 1b system performance for plain entity recognition on the EMEA test corpus. Data shown in *italic font* presents versions of the official runs that were submitted with format corrections after the official deadline. The **official** median and average are computed using the official runs while the *fix* median and average are computed using the late-submission corrected runs

Team	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F-measure**
Erasmus-run1	1720	570	540	0.751	0.761	**0.756**
Erasmus-run2	1753	716	507	0.710	**0.776**	0.741
IHS-RD-run1-fix	*1350*	*223*	*910*	*0.858*	*0.597*	*0.704*
Watchdogs-run1	1238	203	1022	**0.859**	0.548	0.669
IHS-RD-run2-fix	*1288*	*328*	*972*	*0.797*	*0.570*	*0.665*
HIT-WI Lab-run1-fix	*971*	*234*	*1289*	*0.806*	*0.430*	*0.561*
LIMSI-run1	945	644	1315	0.595	0.418	0.491
Watchdogs-run2	1309	2361	951	0.357	0.579	0.442
UPF-run1-fix	*113*	*2147*	*704*	*0,050*	*0,138*	*0,073*
HIT-WI Lab-run1	12	1137	2248	0.010	0.005	0.007
CISMeF-run1	9	4124	2251	0.002	0.004	0.003
IHS-RD-run1	0	0	2260	0.000	0.000	0.000
IHS-RD-run2	0	1616	2260	0.000	0.000	0.000
UPF-run1	0	1067	2260	0.000	0.000	0.000
average (official)				0.328	0.309	0.311
average-fix				0.573	0.468	0.503
median (official)				0.184	0.212	0.224
median-fix				0.731	0.559	0.613

Twelve teams participated in Task 2 with result submissions for the English queries (only one of these teams submitted results for the multilingual queries). On average, teams submitted 8 runs each (the total number of submitted runs by participating teams was 97). Run 3 from Team ECNU performed best under all measures, achieving improvements of up to about 62% and 54% over the best task baseline and the best task benchmark, respectively, and 60% over the second best run from another team. Table 9 summarises the retrieval effectiveness of the best system runs for each participating team and it includes the evaluation results for the most effective task baseline and benchmark systems. Note that average and median system effectiveness are below the task baseline effectiveness, and only five teams achieved results that are more effective than the best task baseline.

Table 4. Task 1b system performance for plain entity recognition on the MEDLINE test corpus. Data shown in *italic font* presents versions of the official runs that were submitted with format corrections after the official deadline. The **official** median and average are computed using the official runs while the *fix* median and average are computed using the late-submission corrected runs

Team	TP	FP	FN	Precision	Recall	F-measure
Erasmus-run1	1861	756	1116	0.711	0.625	**0.665**
Erasmus-run2	1912	886	1065	0.683	**0.642**	0.662
IHS-RD-run1-fix	*1195*	*1782*	*376*	*0.761*	*0.401*	*0.526*
IHS-RD-run2	1188	383	1789	**0.756**	0.399	0.522
Watchdogs-run1	1215	490	1762	0.713	0.408	0.519
LIMSI-run1	1121	834	1856	0.573	0.377	0.455
HIT-WI Lab-run1	1068	671	1909	0.614	0.359	0.453
Watchdogs-run2	1364	2069	1613	0.397	0.458	0.426
CISMeF-run1	680	4412	2297	0.134	0.228	0.169
IHS-RD-run1	75	168	2902	0.309	0.025	0.047
UPF-run1	82	888	2895	0.085	0.028	0.042
average (official)				0.498	0.355	0.396
average-fix				0.543	0.393	0.444
median (official)				0.594	0.388	0.454
median-fix				0.649	0.400	0.487

Table 5. Task 1b system performance for normalized entity recognition on the EMEA test corpus. Data shown in *italic font* presents versions of the official runs that were submitted with format corrections after the official deadline. The **official** median and average are computed using the official runs while the *fix* median and average are computed using the late-submission corrected runs

Team	TP	FP	FN	Precision	Recall	F-measure
CISMeF-run1	10	2255	4128	0.004	0.002	0.003
Erasmus-run1	1637	655	678	**0.714**	**0.707**	**0.711**
Erasmus-run2	1627	680	866	0.705	0.653	0.678
IHS-RD-run1	0	2260	1616	0.000	0.000	0.000
IHS-RD-run1-fix	*923*	*17264*	*710*	*0.051*	*0.565*	*0.093*
HIT-WI Lab-run1	8	2252	1112	0.003	0.007	0.005
HIT-WI Lab-run1-fix	*432*	*1828*	*735*	*0.191*	*0.370*	*0,252*
average (official)				0.286	0.274	0.279
average-fix				0.333	0.460	0.347
median (official)				0.004	0.007	0.005
median-fix				0.191	0.565	0.252

Table 6. Task 1b system performance for normalized entity recognition on the MEDLINE test corpus. Data shown in *italic font* presents versions of the official runs that were submitted with format corrections after the official deadline. The **official** median and average are computed using the official runs while the *fix* median and average are computed using the late-submission corrected runs

Team	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F-measure**
CISMeF-run1	1020	2434	4461	0.295	0.186	0.228
Erasmus-run1	1660	1376	957	**0.547**	**0.634**	**0.587**
Erasmus-run2	1677	1363	1121	0.552	0.599	0.575
IHS-RD-run1	634	15170	938	0.040	0.403	0.073
IHS-RD-run1-fix	*927*	*17495*	*644*	*0.050*	*0.590*	*0.093*
HIT-WI Lab-run1	515	2460	1223	0.173	0.2963	0.219
average (official)				0.321	0.424	0.336
average-fix				0.323	0.461	0.340
median (official)				0.295	0.403	0.228
median-fix				0.295	0.590	0.228

Table 7. Task 1b system performance for entity normalization on the EMEA test corpus

Team	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F-measure**
Erasmus-run1	1734	526	0	0.767	**1.000**	0.868
Erasmus-run2	1748	512	0	**0.774**	**1.000**	**0.872**
IHS-RD-run1	1578	26642	715	0.056	0.688	0.103
HIT-WI Lab-run1	1266	994	1027	0.560	0.552	0.556
average (official)				0.532	0.896	0.615
median (official)				0.767	1.000	0.868

Table 8. Task 1b system performance for entity normalization on the MEDLINE test corpus

Team	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F-measure**
Erasmus-run1	1780	1328	398	0.573	**0.817**	**0.674**
Erasmus-run2	1787	1321	433	**0.575**	0.805	0.671
IHS-RD-run1	1712	38213	1264	0.043	0.575	0.080
HIT-WI Lab-run1	1386	1589	1590	0.466	0.466	0.466
average (official)				0.397	0.733	0.475
median (official)				0.573	0.805	0.671

Table 9. Task 2 system effectiveness. For each participant teams, only the best run (according to p@10) is reported; systems are ranked by p@10. Best retrieval effectiveness are highlighted in bold; task baseline and benchmark effectiveness are reported in italics. Average and median system effectiveness are computed over all (English-only) submitted runs

Run	**p@10**	**nDCG@10**	**RBP**	**uRBP**	**uRBPgr**
ECNU_EN_Run.3	**0.5394**	**0.5086**	**0.5339**	**0.3877**	**0.4046**
KISTI_EN_RUN.6	0.3864	0.3464	0.3332	0.2607	0.2695
CUNI_EN_Run.7	0.3803	0.3465	0.3946	0.3422	0.3312
HCMUS_EN_Run.1	0.3636	0.3323	0.3715	0.3017	0.3062
readability_run.2	*0.3606*	*0.3299*	*0.3756*	*0.3154*	*0.3117*
USST_EN_Run.2	0.3379	0.3000	0.3557	0.2659	0.2727
baseline_run.1	*0.3333*	*0.3151*	*0.3567*	*0.2990*	*0.2933*
Miracl_EN_Run.1	0.3212	0.2787	0.3287	0.2546	0.2631
UBML_EN_Run.2	0.3197	0.2909	0.3305	0.2709	0.2735
GRIUM_EN_Run.6	0.3182	0.2944	0.3306	0.2791	0.2761
YorkU_EN_Run.7	0.3015	0.2766	0.3125	0.2470	0.2523
FDUSGInfo_EN_Run.1	0.2970	0.2718	0.3134	0.2572	0.2568
LIMSI_EN_run.3	0.2621	0.1960	0.2417	0.2036	0.2060
KUCS_EN_Run.1	0.2545	0.2205	0.2785	0.2312	0.2251
average (all runs)	0.2771	0.2529	0.2806	0.2228	0.2247
median (all runs)	0.2970	0.2718	0.3095	0.2394	0.2426

5 Conclusions

In this paper we provided an overview of the third year of the CLEF eHealth evaluation lab. The lab aimed to support the continuum of care by developing methods and resources that make health documents easier to understand, access and author for patients and nurses. Building on the first and second years of the lab, which contained three tasks focusing on IE from clinical reports, information visualization and both mono-lingual and multi-lingual IR, this year's edition featured clinical speech recognition, French IE, and a new mono- and multi-lingual IR challenge. Specifically this year's tasks comprised: 1) Clinical speech recognition related to converting verbal nursing handover to written free-text records; 2) Named entity recognition in clinical reports; and 3) health-focused web search. The lab attracted much interest with 20 teams from around the world submitting a combined total of 174 systems to the shared tasks. Given the significance of the tasks, all test collections and resources associated with the lab have been made available to the wider research community.

Acknowledgement

The CLEF eHealth 2015 evaluation lab has been supported in part by (in alphabetical order) ANR, the French National Research Agency, under grant

CABeRneT ANR-13-JS02-0009-01; NICTA, funded by the Australian Government through the Department of Communications and the Australian Research Council through the Information and Communications Technology (ICT) Centre of Excellence Program; PhysioNetWorks Workspaces; the CLEF Initiative; and the Khresmoi project, funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 257528. We express our gratitude to Maricel Angel, Registered Nurse at NICTA, for helping us to create the Task 1a dataset, to Dr. Johannes Bernhardt-Melischnig (Medizinische Universität Graz) for coordinating the recruitment and management of the paid medical students that participated in the relevance assessment exercise, and to the volunteering students and researchers at the organisers' universities for participating to the query generation 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. 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. Springer Berlin Heidelberg (2013) 212–231
2. Pradhan, S., Elhadad, N., South, B., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W., Savova, G.: Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. In: Online Working Notes of CLEF, CLEF (2013)
3. Mowery, D., South, B., Christensen, L., Murtola, L., Salanterä, S., Suominen, H., Martinez, D., Elhadad, N., Pradhan, S., Savova, G., Chapman, W.: Task 2: ShARe/CLEF eHealth Evaluation Lab 2013. In: Online Working Notes of CLEF, CLEF (2013)
4. Goeuriot, L., Jones, G., 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. In: Online Working Notes of CLEF, CLEF (2013)
5. Kelly, L., Goeuriot, L., 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. Springer Berlin Heidelberg (2014) 172–191
6. Suominen, H., Schreck, T., Leroy, G., Hochheiser, H., Goeuriot, L., Kelly, L., Mowery, D., Nualart, J., Ferraro, G., Keim, D.: Task 1 of the CLEF eHealth Evaluation Lab 2014: visual-interactive search and exploration of eHealth data. In: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes, Sheffield, UK (2014)
7. Mowery, D., Velupillai, S., South, B., Christensen, L., Martinez, D., Kelly, L., Goeuriot, L., Elhadad, N., Pradhan, S., Savova, G., Chapman, W.: Task 2 of the CLEF eHealth Evaluation Lab 2014: Information extraction from clinical text. In: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes, Sheffield, UK (2014)
8. 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)

9. Suominen, H., Hanlen, L., Goeuriot, L., Kelly, L., Jones, G.J.: Task 1a of the CLEF eHealth evaluation lab 2015: Clinical speech recognition. In: CLEF 2015 Online Working Notes, CEUR-WS (2015)
10. 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)
11. 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)
12. Névéol, A., Dalianis, H., Savova, G., Zweigenbaum, P.: Didactic panel: Clinical natural language processing in languages other than English. In: Proc AMIA Annu Symp. (2014)
13. Goeuriot, L., Kelly, L., Jones, G.J., Zuccon, G., Suominen, H., Hanbury, A., Mueller, H., Leveling, J.: Creation of a new evaluation benchmark for information retrieval targeting patient information needs. In: The Fifth International Workshop on Evaluating Information Access (EVIA 2013). Volume 18., Dublin City University (2013)
14. Fox, S.: Health topics: 80% of internet users look for health information online. Pew Internet & American Life Project (2011)
15. Benigeri, M., Pluye, P.: Shortcomings of health information on the internet. Health promotion international **18**(4) (2003) 381–386
16. White, R.W., Horvitz, E.: Cyberchondria: studies of the escalation of medical concerns in web search. ACM TOIS **27**(4) (2009) 23
17. Zuccon, G., Koopman, B., Palotti, J.: Diagnose this if you can: On the effectiveness of search engines in finding medical self-diagnosis information. In: Advances in Information Retrieval. Springer (2015) 562–567
18. Suominen, H., Zhou, L., Hanlen, L., Ferraro, G.: Benchmarking clinical speech recognition and information extraction: New data, methods and evaluations. JMIR Medical Informatics **3**(2) (2015) e19
19. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: MIE village of the future. (2012)
20. Bodenreider, O., McCray, A.T.: Exploring semantic groups through visual approaches. J Biomed Inform **36**(6) (2003) 414–32
21. Névéol, A., Grouin, C., Leixa, J., Rosset, S., Zweigenbaum, P.: The QUAERO French medical corpus: A ressource for medical entity recognition and normalization. In: Proc of BioTextMining Work. (2014) 24–30
22. Stanton, I., Ieong, S., Mishra, N.: Circumlocution in diagnostic medical queries. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval, ACM (2014) 133–142
23. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6) (1945) 80–83
24. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4) (2002) 422–446
25. Zuccon, G., Koopman, B.: Integrating understandability in the evaluation of consumer health search engines. In: Medical Information Retrieval Workshop at SIGIR 2014. (2014) 32
26. Verspoor, K., Yepes, A.J., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., Plazzer, J.P.: Annotating the biomedical literature for the human variome. Database (Oxford) (2013) bat019–bat019