Overview of the ShARe/CLEF eHealth Evaluation Lab 2014

Liadh Kelly[1], Lorraine Goeuriot[1], Hanna Suominen[2], Tobias Schreck[3], Gondy Leroy[4], Danielle L. Mowery[5], Sumithra Velupillai[6], Wendy W. Chapman[7], David Martinez[8], Guido Zuccon[9], and Joao Palotti[10] *

[1] Dublin City University, Ireland, `Firstname.Lastname@computing.dcu.ie`
[2] NICTA, The Australian National University, University of Canberra, and University of Turku, ACT, Australia, `Hanna.Suominen@nicta.com.au`
[3] University of Konstanz, Germany, `tobias.schreck@uni-konstanz.de`
[4] University of Arizona, Tucson, AZ, USA, `gondyleroy@email.arizona.edu`
[5] University of Pittsburgh, Pittsburgh, Pennsylvania, United States `dlm31@pitt.edu`
[6] Stockholm University, Sweden, `sumithra@dsv.su.se`
[7] University of Utah, Salt Lake City, Utah, United States `wendy.chapman@utah.edu`
[8] University of Melbourne, VIC, Australia `david.martinez@nicta.com.au`
[9] Queensland University of Technology, Australia `g.zuccon@qut.edu.au`
[10] Vienna University of Technology, Austria `palotti@ifs.tuwien.ac.at`

Abstract. This paper reports on the 2nd ShARe/CLEFeHealth evaluation lab which continues our evaluation resource building activities for the medical domain. In this lab we focus on patients' information needs as opposed to the more common campaign focus of the specialised information needs of physicians and other healthcare workers. The usage scenario of the lab is to ease patients and next-of-kins' ease in understanding eHealth information, in particular clinical reports. The 1st ShARe/CLEFeHealth evaluation lab was held in 2013. This lab consisted of three tasks. Task 1 focused on named entity recognition and normalization of disorders; Task 2 on normalization of acronyms/abbreviations; and Task 3 on information retrieval to address questions patients may have when reading clinical reports. This year's lab introduces a new challenge in Task 1 on visual-interactive search and exploration of eHealth data. Its aim is to help patients (or their next-of-kin) in readability issues related to their hospital discharge documents and related information search on the Internet. Task 2 then continues the information extraction work of the 2013 lab, specifically focusing on disorder attribute identification and normalization from clinical text. Finally, this year's Task 3 further extends the 2013 information retrieval task, by cleaning the 2013 document collection and introducing a new query generation method and multilingual queries. De-identified clinical reports used by the three tasks were from US intensive care and originated from the MIMIC II database. Other text documents for Tasks 1 and 3 were from the Internet and originated from the Khresmoi project. Task 2 annotations originated from

* In alphabetical order, LK & LG co-chaired the lab & led Task 3; DLM, SV & WWC led Task 2; and DM, GZ & JP were the leaders of result evaluations. In order of contribution HS, TS & GL led Task 1.

the ShARe annotations. For Tasks 1 and 3, new annotations, queries, and relevance assessments were created. 50, 79, and 91 people registered their interest in Tasks 1, 2, and 3, respectively. 24 unique teams participated with 1, 10, and 14 teams in Tasks 1, 2 and 3, respectively. The teams were from Africa, Asia, Canada, Europe, and North America. The Task 1 submission, reviewed by 5 expert peers, related to the task evaluation category of Effective use of interaction and targeted the needs of both expert and novice users. The best system had an Accuracy of 0.868 in Task 2a, an F1-score of 0.576 in Task 2b, and Precision at 10 (P@10) of 0.756 in Task 3. The results demonstrate the substantial community interest and capabilities of these systems in making clinical reports easier to understand for patients. The organisers have made data and tools available for future research and development.

Keywords: Information Retrieval, Information Extraction, Information Visualisation, Evaluation, Medical Informatics, Test-set Generation, Text Classification, Text Segmentation

1 Introduction

Laypeople find eHealth clinical reports, such as discharge summaries and radiology reports, difficult to understand. Clinicians also experience difficulties in understanding the jargon of other professional groups even though laws and policies emphasise patients' right to be able to access and understand their clinical documents. A simple example from a US discharge document is "*AP: 72 yo f w/ ESRD on HD, CAD, HTN, asthma p/w significant hyperkalemia & associated arrythmias*". As described in [1], there is much need for techniques which support individuals in understanding such eHealth documents.

The usage scenario of the CLEF eHealth lab is to ease patients and next-of-kins' ease in understanding eHealth information. eHealth documents are much easier to understand after expanding shorthand, correcting misspellings and normalising all health conditions to standardised terminology. This would result in "*Description of the patient's active problem: 72 year old female with dependence on hemodialysis, coronary heart disease, hypertensive disease, and asthma who is currently presenting with the problem of significant hyperkalemia and associated arrhythmias.*" The patient's and her next-of-kin's understanding of health conditions can also be supported by linking discharge summary terms to a patient-centric search on the Internet. The search engine could, for example, link hyperkalemia and its synonyms to definitions in Wikipedia, Consumer Health Vocabulary, and other patient-friendly sources[11]. This would explain the connection between hyperkalemia and arrhythmia: *Extreme hyperkalemia (having too much potassium in the blood) is a medical emergency due to the risk of potentially fatal arrhythmias (abnormal heart rhythms).* The engine should

[11] `http://en.wikipedia.org/` and `http://www.consumerhealthvocab.org/`

also assess the reliability of information (e.g., guidelines by healthcare service providers vs. uncurated but insightful experiences on discussion forums).

Natural language processing (NLP), computational linguistics and machine learning are recognised as ways to process textual health information. Several evaluation campaigns have been organised to share benchmarks and improve techniques such as information retrieval (IR), text mining, image retrieval and processing, etc. We described these campaigns in detail in [1].

This paper presents an overview of the ShARe/CLEFeHealth2014 evaluation lab[12] to support development of approaches which support patients' and their next-of-kins' information needs stemming from clinical reports. Towards this, this second year of the novel lab aimed to build on the resource building and evaluation approaches offered by the first year of the lab. The first year of the lab contained two tasks which focused on named entity recognition and normalization of disorders and acronyms/abbreviations in clinical reports [2, 3], and one task which explored supporting individuals' information needs stemming from clinical reports through IR technique development [4]. This years' lab expands our year one efforts and supports evaluation of information visualisation (Task 1), information extraction (Task 2) and information retrieval (Task 3) approaches for the space. Specifically, Task 1 [5] aims to help patients (or their next-of-kin) in readability issues related to their hospital discharge documents and related information search on the Internet. Task 2 [6] continues the information extraction work of the 2013 CLEFeHealth lab, specifically focusing on information extraction of disorder attributes from clinical text. Task 3 [7] further extends the 2013 information retrieval task, by cleaning the 2013 document collection and introducing a new query generation method and multilingual queries.

In total the 2014 edition of the CLEFeHealth lab attracted 24 teams to submit 105 systems[13]; demonstrated the capabilities of these systems in contributing to patients' 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 2014.

2 Materials and Methods

2.1 Text Documents

For Tasks 2 and 3, de-identified clinical reports were from US intensive care and originated from the ShARe corpus which has added layers of annotation over the clinical notes in the version 2.5 of the MIMIC II database[14]. The corpus consisted of discharge summaries, electrocardiogram, echocardiogram, and

[12] `http://clefehealth2014.dcu.ie/`, Shared Annotated Resources, `http://clinicalnlpannotation.org`, and Conference and Labs of the Evaluation Forum, `http://www.clef-initiative.eu/`

[13] Note: in this paper we refer to systems, experiments, and runs as *systems*.

[14] Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5, `http://mimic.physionet.org`

radiology reports. They were authored in the intensive care setting. Although the clinical reports were de-identified, they still needed to be treated with appropriate care and respect. Hence, all participants were required to register to the lab, obtain a US human subjects training certificate[15], create an account to a password-protected site on the Internet, specify the purpose of data usage, accept the data use agreement, and get their account approved. Six of these clinical reports were further de-identified for use in Task 1. This was done by organisers manually removing any remaining potentially identifying information, e.g. treatment hospital, from the reports.

For Tasks 1 and 3, an updated version of the CLEFeHealth 2013 Task 3 large crawl of health resources on the Internet was used. In this updated crawl, the 2013 Task 3 crawl was further cleaned, by removing some errors in HTML, duplicate documents, etc. It contained about one million documents [8] and originated from the Khresmoi project[16]. 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.[17] 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. [18]

2.2 Human Annotations, Queries, and Relevance Assessments

For Task 1 the input data provided to participants consists of *six carefully chosen cases* from the CLEFeHealth2013 data set. Using the first case was mandatory for all participants and the other five cases were optional. Each case consisted of a discharge summary, including the disease/disorder spans marked and mapped to *Systematized Nomenclature of Medicine Clinical Terms, Concept Unique Identifiers* (SNOMED-CT), and the shorthand spans marked and mapped to the *Unified Medical Language System* (UMLS). Each discharge summary was also associated with a *profile* to describe the patient, a *narrative* to describe her information need, a *query* to address this information need by searching the Internet documents, and the list of *returned relevant documents*. To access the data set on the *PhysioNetWorks workspaces*, the participants had to first register to CLEF2014 and agree to our data use agreement. The dataset was accessible

[15] The course was available free of charge on the Internet, for example, via the CITI Collaborative Institutional Training Initiative at https://www.citiprogram.org/Default.asp or the US National Institutes of Health (NIH) at http://phrp.nihtraining.com/users/login.php.

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

[17] 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

[18] HyperText Markup Language and Uniform Resource Locators

to authorized users from December 2013. The data set is to be opened for all registered PhysioNetWorks users in October 2014.

For Task 2, the annotations were created as part of the ongoing Shared Annotated Resources (ShARe) project. For this year's evaluation lab, the annotations extended the existing disorder annotations from clinical text from Task 1 ShARe/CLEF eHealth 2013 by focusing on template filling for each disorder mention[19]. As such, each disorder template consisted of 10 different attributes including *Negation Indicator*, *Subject Class*, *Uncertainty Indicator*, *Course Class*, *Severity Class*, *Conditional Class*, *Generic Class*, *Body Location*, *DocTime Class*, and *Temporal Expression*. Each attribute contained two types of annotation values: normalization and cue detection value with the exception of the *DocTime Class* which did not contain a cue detection value. Each note was annotated by two professional coders trained for this task, followed by an open adjudication step. The initial development set contained 300 documents of 4 clinical report types - discharge summaries, radiology, electrocardiograms, and echocardiograms. The unseen test set contained 133 documents of only discharge summaries.

From the ShARe guidelines, for a <u>disorder mention</u>, an **attribute** *cue* is a span of text that represents a non-default normalization value (*default normalization value):

Negation Indicator: def. indicates a disorder was negated: *no, *yes*
Ex. *No* <u>cough</u>.

Subject Class: def. indicates who experienced a disorder: *patient, *family_ member*, donor_family_member, donor_other, null, other
Ex. *Dad* had <u>MI</u>.

Uncertainty Indicator: def. indicates a measure of doubt about the disorder: *no, *yes*
Ex. *Possible* <u>pneumonia</u>.

Course Class: def. indicates progress or decline of a disorder: *unmarked, changed, increased, decreased, improved, worsened, *resolved*
Ex. <u>Bleeding</u> *abated.*

Severity Class: def. indicates how severe a disorder is: *unmarked, slight, moderate, *severe*
Ex. <u>Infection</u> is *severe.*

Conditional Class: def. indicates existence of disorder under certain circumstances: *false, *true*
Ex. Return *if* <u>nausea</u> occurs.

[19] `http://clefehealth2014.dcu.ie/task-2/2014-dataset`

Generic Class: def. indicates a generic mention of disorder: *false, *true*
Ex. Vertigo *while* walking.

Body Location: def. represents an anatomical location: *NULL, *CUI: C0015450*, CUI-less
Ex. *Facial* lesions.

DocTime Class: def. indicates temporal relation between a disorder and document authoring time: *before*, after, overlap, before-overlap, *unknown
Ex. Stroke in *1999*.

Temporal Expression: def. represents any TIMEX (TimeML) temporal expression related to the disorder: *none, *date*, time, duration, set
Ex. Flu on *March 10*.

For Task 3, queries and the respective result sets were associated with the text documents. Two Finnish nursing professionals created 55 queries from the main disorders diagnosed in discharge summaries provided in Task 1 (semi-automatically identified). Participants were provided with the mapping between queries and discharge summaries, and were free to use the discharge summaries. Relevance assessments were performed by domain experts and technological experts using the Relevation system[20] [9] for collecting relevance assessments of documents contained in the assessment pools. Documents and queries were uploaded to the system via a browser-based interface; judges could browse documents for each query and provide their relevance judgements. The domain experts included two Indian medical professionals, and two Finnish nursing professionals. The technological experts included six Irish, five Czech, one Austrian and one Australian senior researcher in clinical NLP and machine learning (ML). Assessments compared the query and its mapping to the content of the retrieved document on a four-point scale. These graded relevance assessments yielded 0: 3,044, 1: 547, 2: 974, 3: 2,235 documents. The relevance of each document was assessed by one expert. The 55 queries were divided into 5 training and 50 test queries. Assessments for the 5 training queries were performed by the same two Finnish nursing professionals who generated the queries. As we received 65 systems, we had to limit the pool depth for the test set of 50 queries and distribute the relevance assessment workload between domain experts and technological experts. System outputs for 35 test queries were assessed by the domain experts and the remaining 15 test queries by the technological experts.

2.3 Evaluation Methods

The following evaluation criteria were used: In Task 1, each final submission was assessed by a team of four evaluation panellists, supported by an orga-

[20] `https://github.com/bevankoopman/relevation`, open source, based on Python's Django Internet framework, uses a simple Model-View-Controller model that is designed for easy customisation and extension

nizer. Primary evaluation criteria included the effectiveness and originality of the presented submissions. More precisely, submissions were judged on usability, visualization, interaction, and aesthetics. In Task 2 evaluation was based on correctness in assigning normalization values to ten semantic attributes attributes (2a), and correctness in assigning cue values to the nine semantic attributes with cues (2b), and in Task 3 relevance of the retrieved documents to patients or their representatives based on English queries (3a) or non-English queries translated into English (3b).

In Task 1, teams were asked to submit the following mandatory items by 1 May 2014:

1. a concise report of the design, implementation (if applicable), and application results discussion in the form of an extended abstract that highlights the obtained findings, possibly supported by an informal user study or other means of validation and
2. two demonstration videos illustrating the relevant functionality of the functional design or paper prototype in application to the provided task data.

In the first video, the user should be a person who knows the system functionalities and in the second video, the user should be a novice with no previous experience of these functionalities. The video should also explain how the novice was trained to use the functionality.

In Tasks 2a and 2b, each participating team was permitted to upload the outputs of up to two systems. Task 2b was optional for Task 2 participants. In Task 3a, teams were asked to submit up to seven ranked outputs (typically called *runs*): a mandatory baseline (referred to as {team}.run1): only title and description in the query could be used without any additional resources (e.g., clinical reports, corpora, or ontologies); up to three outputs from systems which use the clinical reports (referred to as {team}.run2–{team}.run4); and up to three outputs from systems which do not use the clinical reports (referred to as {team}.run5–{team}.run7). One of the runs 2–4 and one of the runs 5–7 needed to use only the fields title and description from the queries. The ranking corresponded to priority (referred to as {team}.{run}.{rank} with ranks 1–7 from the highest to lowest priority). In Task 3b, teams could submit a similar set of ranked outputs for each of the cross-lingual languages.

Teams received data from December 2013 to April 2014. In Task 1, all data was accessible to authorized users from December, 2013. In Tasks 2 and 3, data was divided into training and test sets; the evaluation for these tasks was conducted using the blind, withheld test data (reports for Task 2 and queries for Task 3). Teams were asked to stop development as soon as they downloaded the test data. The training set and test set for Tasks 2 and 3 were released from December 2013 and April 2014 respectively. Evaluation results were announced to the participants for the three tasks from end May to early June.

In Tasks 2a and 2b, participants were provided with a training set containing clinical text as well as pre-annotated spans and CUIs for diseases/disorders in templates along with 1) normalized values for each of the ten attributes of the disease/disorder (Task 2a) and cue slot values for nine of the attributes (Task

2b). For Task 2a, participants were instructed to develop a system that kept or updated the normalization values for the ten attributes. For Task 2b, participants were instructed to develop a system that kept or updated the cue values for the nine attributes. The outputs needed to follow the annotation format. The corpus of reports was split into 300 training and 133 testing.

In Task 3, post-submission relevance assessment of systems trained on the 5 training queries and the matching result set was conducted on the 50 test queries to generate the complete result set. The outputs needed to follow the TREC format. The top ten documents obtained from the participants' baseline, the two highest priority runs from the runs 2–4, and the two highest priority output from the runs 5–7[21] were pooled with duplicates removed. This resulted in a pool of 6,040 documents, with a total of 6,800 relevance judgements.[22] Pooled sets for the training queries were created by merging the top 30 ranked documents returned by the two IR models (Vector Space Model [10] and BM25 [11]) and removing duplicates.

The system performance in the different tasks was evaluated against task-specific criteria. Task 1 aimed at providing a visual-interactive application to help users explore data and understand complex relationships. As such, an evaluation in principle needs to consider multiple dimensions regarding the system design, including effectiveness and expressiveness of the chosen visual design, and criteria of usability by different user groups. Specifically, in Task 1 participants were asked to demonstrate that their design addresses the posed user tasks, gives a compelling use-case driven discussions, and highlight obtained findings. Furthermore, we devised a set of usability and visualization heuristics to characterize the quality of the solution.

Tasks 2 and 3 system performance was evaluated using Accuracy in Task 2a and the F1-score in Task 2b, and Precision at 10 (P@10) and Normalised Discounted Cumulative Gain at 10 (NDCG@10) in Task 3. We relied on the Wilcoxon test [12] in Task 3 to better compare the measure values for the systems and benchmarks.

In Task 2a, the Accuracy was defined as the number of correctly predicted normalization value slots divided by the total number of gold standard normalization slot values.

In Task 2b, the F1 score was defined as the harmonic mean of Precision (P) and Recall (R); P as $n_{TP}/(n_{TP} + n_{FP})$; R as $n_{TP}/(n_{TP} + n_{FN})$; n_{TP} as the number of instances, where the spans identified by the system and gold standard were the same; n_{FP} as the number of spurious spans by the system; and n_{FN} as the number of missing spans by the system. We referred to the Exact (Relaxed) F1-score if the system span is identical to (overlaps) the gold standard span.

In Task 2b, the Exact F1-score and Relaxed F1-score were measured. In the Exact F1-score for Task 2b, the predicted cue slot span was identical to the reference standard span. In the Relaxed F1-score, the predicted cue slot span overlapped with reference standard span.

[21] Runs 1, 2, 3, 5 and 6 for teams who submitted the maximum number of runs.
[22] This means that some documents have been retrieved for several queries.

In Task 3, the official primary and secondary measures were P@10 and NDCG@10 [13], respectively. Both measures were calculated over the top ten documents retrieved by a system for each query, and then averaged across the whole set of queries. To compute P@10, graded relevance assessments were converted to a binary scale; NDCG@10 was computed using the original relevance assessments on a 4-point scale. The `trec_eval` evaluation tool[23] was used to calculate these evaluation measures[24]. Participants were also provided with other standard measures calculated by `trec_eval`[25].

The organisers provided the following evaluation tools on the Internet: a evaluation script for calculation of the evaluation measures of Task 2; a Graphical User Interface (GUI) for visualisation of gold standard annotations; and a pointer to the `trec_eval` evaluation tool for Task 3.

3 Results

The number of people who registered their interest in Tasks 1, 2, and 3 was 50, 79, and 91, respectively, and in total 24 teams with unique affiliations submitted to the shared tasks (Table 1). No team participated in all three tasks. One team participated in Tasks 2 and 3 (Table 2). Teams represented Canada, Czech Republic, France, Germany, India, Japan, Portugal, Spain, South Korea, Taiwan, Thailand, The Netherlands, Tunisia, Turkey, Vietnam, and USA.

In total 105 systems were submitted to the challenge (Table 2).

In Task 1, one final submission was received from a team from the USA called *FLPolytech*. This submission was also assessed during our optional draft submission round in March 2014. The team was a partnership between *Florida Polytechnic University's Department of Advanced Technology* and the commercial information science firm *Retrivika*. The submission addressed both Tasks *1a: Discharge Resolution Challenge* and *1b: Visual Exploration Challenge* together with their integration as the *Grand Challenge* solution. It related to the task evaluation category of *Effective use of interaction*. Although the submission did not describe tests with real expert and/or novice users, the described system appeared to be rather good. The final submission was evaluated by four evaluation panellists and one organizer. The draft submission was reviewed by five organizers.

In total, ten teams submitted systems for Task 2a. Four teams submitted two runs. For Task 2b, three teams submitted systems, one of them submitted two runs. See Table 2. The best system had an Accuracy of 0.868 in Task 2a and an F1-score of 0.576 in Task 2b. See Tables 3 - 6 for details.

Fourteen teams participated in Task 3a. Two of these teams also participated in Task 3b. The number of submissions per team ranged from 1-7. See Table 2.

[23] `http://trec.nist.gov/trec_eval/`

[24] NDCG was computed with the standard settings in `trec_eval`, and by running the command `trec_eval -c -M1000 -m ndcg_cut qrels runName`.

[25] including P@5, NDCG@5, Mean Average Precision (MAP), and rel_ret (i.e., the total number of relevant documents retrieved by the system over all queries)

The best system in Task 3a had P@10 of 0.756 and NDCG@10 of 0.7445; and the best system in Task 3b had P@10 of 0.7551 and NDCG@10 of 0.7011. See Tables 7 - 9 for details.

4 Conclusions

In this paper we provided an overview of the second year of the ShARe/CLEF eHealth evaluation lab. The lab aims to support the continuum of care by developing methods and resources that make clinical reports and related medical conditions easier to understand for patients. The focus on patients' information needs as opposed to the specialised information needs of healthcare workers is the main distinguishing feature of the lab from previous shared tasks on NLP, ML and IR in the space. Building on the first year of the lab which contained three tasks focusing on information extraction from clinical reports and a mono-lingual information retrieval, this years edition featured an information visualisation challenge, further information extraction challenges and multi-lingual information retrieval. Specifically this year's three tasks comprised: 1) Visual-Interactive Search and Exploration of eHealth Data; 2) Information extraction from clinical text; and 3) User-centred health information retrieval. The lab attracted much interest with 24 teams from around the world submitting a combined total of 105 systems to the shared tasks. Given the significance of the tasks, all test collections, etc associated with the lab have been made available to the wider research community.

Acknowledgement

The ShARe/CLEF eHealth 2014 evaluation lab has been supported in part by (in alphabetical order) MIMIC II Database; NICTA, funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program; PhysioNetWorks Workspaces; the CLEF Initiative; the Khresmoi project, funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 257528; the ShARe project funded by the United States National Institutes of Health (R01GM090187); the US Office of the National Coordinator of Healthcare Technology, Strategic Health IT Advanced Research Projects (SHARP) 90TR0002; and the Swedish Research Council (350-2012-6658).

We acknowledge the generous support of time and expertise that the evaluation panelists (Chih-Hao (Justin) Ku, Assistant Professor in Text mining and information visualization, Lawrence Technological University, Southfield, MI, USA; Hilary Cinis, Senior User Experience Designer, NICTA, Sydney, NSW, Australia; Lin Shao, PhD student, in Computer and Information Science, University of Konstanz, Konstanz, Germany; and Mitchell Whitelaw, Associate Professor in Media Arts and Production, University of Canberra, Canberra ACT, Australia), annotators as well as members of the organising and mentoring committees have invested in this evaluation lab. We also acknowledge the contribution of George Moody, Harvard-MIT, Cambridge, MA, USA in proofing and supporting the release of our six double de-identified (manually and automatically) discharge summaries.

Table 1. Participating teams.

ID	Team	Affiliation	Location
1	ASNLP	iis, sinica	Taiwan
2	CORAL	University of Alabama at Birmingham	USA
3	CSKU/COMPL	Kasetsart University - Department of Computer Science	Thailand
4	CUNI	Charles University in Prague	Czech Republic
5	DEMIR	DEMIR-Dokuz Eylul University, Multimedia Information Retrieval Group	Turkey
6	DFKI-Medical	DFKI	Germany
7	ERIAS	ISPED/ UniversitÃľ of Bordeaux	France
8	FLPolytech	Florida Polytechnic University'd Department of Advanced Technology and Retrivika	USA
9	GRIUM	Departement of Computer Science and Operations Research, University of Montreal	Canada
10	HCMUS	HCM City University of Science	Vietnam
11	HITACHI	Research and Development Centre, Hitachi India Pvt Ltd, Hitachi, Ltd., Central Research Laboratory, Japan, International Institute of Information Technology Hyderabad, India	India, Japan
12	HPI	Hasso Plattner Institute	Germany

ID	Team	Affiliation	Location
13	IRLabDAIICT	DAIICT	India
14	KISTI	Korea Institute of Science and Technology Information	South Korea
15	LIMSI	LIMSI-CNRS	France
16	Miracl	Multimedia Information Systems and Advanced Computing Laboratory	Tunisia
17	Nijmegen	Information Foraging Lab, Institute for Computing and Information Sciences	The Netherlands
18	RelAgent	RelAgent Tech Pvt Ltd	India
19	RePaLi	Inria - IRISA - CNRS	France
20	SNUMEDINFO	Seoul National University	South Korea
21	UEvora	Universidade de ÃĽvora	Portugal
22	UHU	Universidad de Huelva	Spain
23	UIOWA	The University of Iowa	USA
24	YORKU	York University	Canada

Table 2. The tasks that the teams participated in.

ID	Team	Number of submitted systems per task					
		1	2a	2b	3a	3b	
1	ASNLP		1				
2	CORAL		1				
3	CSKU/COMPL				2		
4	CUNI				4	4 runs/language	
5	DEMIR				4		
6	DFKI-Medical		2				
7	ERIAS				4		
8	FLPolytech	1					
9	GRIUM		1		4		
10	HCMUS		1	1			
11	HITACHI		2	2			
12	HPI		1	1			
13	IRLabDAIICT				6		
14	KISTI				7		
15	LIMSI		2				
16	Miracl				1		
17	Nijmegen				7		
18	RelAgent		2				
19	RePaLi				4		
20	SNUMEDINFO				7	4 runs/language	
21	UEvora		1				
22	UHU				4		
23	UIOWA				4		
24	YORKU				4		
	Systems:	1	14	4	62	24	*Total: 105*
	Teams:	1	10	3	14	2	

Table 3. Evaluation in Task 2a: predict each attribute's normalization slot value. Accuracy: overall

Attribute	System ID ({team}.{system})	Accuracy
Overall	TeamHITACHI.2	0.868
Average	TeamHITACHI.1	0.854
	RelAgent.2	0.843
	RelAgent.1	0.843
	TeamHCMUS.1	0.827
	DFKI-Medical.2	0.822
	LIMSI.1	0.804
	DFKI-Medical.1	0.804
	TeamUEvora.1	0.802
	LIMSI.2	0.801
	ASNLP.1	0.793
	TeamCORAL.1	0.790
	TeamGRIUM.1	0.780
	HPI.1	0.769

Table 4. Evaluation in Task 2a: predict each attribute's normalization slot value. Accuracy per attribute type - Attributes Negation Indicator, Subject Class, Uncertainty Indicator, Course Class, Severity Class, Conditional Class.

Attribute	System ID	Accuracy	Attribute	System ID	Accuracy
Negation Indicator	TeamHITACHI.2	0.969	Subject Class	TeamHCMUS.1	0.995
	RelAgent.2	0.944		TeamHITACHI.2	0.993
	RelAgent.1	0.941		TeamHITACHI.1	0.990
	TeamASNLP	0.923		TeamUEvora.1	0.987
	TeamGRIUM.1	0.922		DFKI-Medical.1	0.985
	TeamHCMUS.1	0.910		DFKI-Medical.2	0.985
	LIMSI.1	0.902		LIMSI.1	0.984
	LIMSI.2	0.902		RelAgent.2	0.984
	TeamUEvora.1	0.901		RelAgent.1	0.984
	TeamHITACHI.1	0.883		LIMSI.2	0.984
	DFKI-Medical.2	0.879		TeamHPI	0.976
	DFKI-Medical.1	0.876		TeamCORAL.1	0.926
	TeamCORAL.1	0.807		TeamASNLP	0.921
	TeamHPI	0.762		TeamGRIUM.1	0.611
Uncertainty Indicator	TeamHITACHI.1	0.960	Course Class	TeamHITACHI.2	0.971
	RelAgent.2	0.955		TeamHITACHI.1	0.971
	RelAgent.1	0.955		RelAgent.1	0.970
	TeamUEvora.1	0.955		RelAgent.2	0.967
	TeamCORAL.1	0.941		TeamGRIUM.1	0.961
	DFKI-Medical.1	0.941		TeamCORAL.1	0.961
	DFKI-Medical.2	0.941		TeamASNLP	0.953
	TeamHITACHI.2	0.924		TeamHCMUS.1	0.937
	TeamGRIUM.1	0.923		DFKI-Medical.1	0.932
	TeamASNLP	0.912		DFKI-Medical.2	0.932
	TeamHPI	0.906		TeamHPI	0.899
	TeamHCMUS.1	0.877		TeamUEvora.1	0.859
	LIMSI.1	0.801		LIMSI.1	0.853
	LIMSI.2	0.801		LIMSI.2	0.853
Severity Class	TeamHITACHI.2	0.982	Conditional Class	TeamHITACHI.1	0.978
	TeamHITACHI.1	0.982		TeamUEvora.1	0.975
	RelAgent.2	0.975		RelAgent.2	0.963
	RelAgent.1	0.975		RelAgent.1	0.963
	TeamGRIUM.1	0.969		TeamHITACHI.2	0.954
	TeamHCMUS.1	0.961		TeamGRIUM.1	0.936
	DFKI-Medical.1	0.957		LIMSI.1	0.936
	DFKI-Medical.2	0.957		TeamASNLP	0.936
	TeamCORAL.1	0.942		LIMSI.2	0.936
	TeamUEvora.1	0.919		TeamCORAL.1	0.936
	TeamHPI	0.914		DFKI-Medical.1	0.936
	TeamASNLP	0.912		DFKI-Medical.2	0.936
	LIMSI.1	0.900		TeamHCMUS.1	0.899
	LIMSI.2	0.900		TeamHPI	0.819

Table 5. Evaluation in Task 2a: predict each attribute's normalization slot value. Accuracy per attribute type - Attributes Generic Class, Body Location, DocTime Class and Temporal Expression.

Attribute	System ID	Accuracy	Attribute	System ID	Accuracy
Generic Class	TeamGRIUM.1	1.000	Body Location	TeamHITACHI.2	0.797
	LIMSI.1	1.000		TeamHITACHI.1	0.790
	TeamHPI	1.000		RelAgent.2	0.756
	TeamHCMUS.1	1.000		RelAgent.1	0.753
	RelAgent.2	1.000		TeamGRIUM.1	0.635
	TeamASNLP	1.000		DFKI-Medical.2	0.586
	RelAgent.1	1.000		TeamHCMUS.1	0.551
	LIMSI.2	1.000		TeamASNLP	0.546
	TeamUEvora.1	1.000		TeamCORAL.1	0.546
	DFKI-Medical.1	1.000		TeamUEvora.1	0.540
	DFKI-Medical.2	1.000		LIMSI.1	0.504
	TeamHITACHI.2	0.990		LIMSI.2	0.504
	TeamCORAL.1	0.974		TeamHPI	0.494
	TeamHITACHI.1	0.895		DFKI-Medical.1	0.486
DocTime Class	TeamHITACHI.2	0.328	Temporal Expression	TeamHPI	0.864
	TeamHITACHI.1	0.324		RelAgent.2	0.864
	LIMSI.1	0.322		RelAgent.1	0.864
	LIMSI.2	0.322		TeamCORAL.1	0.864
	TeamHCMUS.1	0.306		TeamUEvora.1	0.857
	DFKI-Medical.1	0.179		DFKI-Medical.2	0.849
	DFKI-Medical.2	0.154		LIMSI.1	0.839
	TeamHPI	0.060		TeamHCMUS.1	0.830
	TeamGRIUM.1	0.024		TeamASNLP	0.828
	RelAgent.2	0.024		TeamGRIUM.1	0.824
	RelAgent.1	0.024		LIMSI.2	0.806
	TeamUEvora.1	0.024		TeamHITACHI.2	0.773
	TeamASNLP	0.001		TeamHITACHI.1	0.766
	TeamCORAL.1	0.001		DFKI-Medical.1	0.750

Table 6. Evaluation in Task 2b: predict each attribute's cue slot value. Strict and Relaxed F1-score, Precision and Recall (overall and per attribute type)

Attribute	System ID	Strict			Relaxed		
		F1-score	Precision	Recall	F1-score	Precision	Recall
Overall Average	TeamHITACHI.2	0.676	0.620	0.743	0.724	0.672	0.784
	TeamHITACHI.1	0.671	0.620	0.731	0.719	0.672	0.773
	TeamHCMUS.1	0.544	0.475	0.635	0.648	0.583	0.729
	HPI.1	0.190	0.184	0.197	0.323	0.314	0.332
Negation Indicator	TeamHITACHI.2	0.913	0.955	0.874	0.926	0.962	0.893
	TeamHITACHI.1	0.888	0.897	0.879	0.905	0.912	0.897
	TeamHCMUS.1	0.772	0.679	0.896	0.817	0.735	0.919
	HPI.1	0.383	0.405	0.363	0.465	0.488	0.444
Subject Class	TeamHCMUS.1	0.857	0.923	0.800	0.936	0.967	0.907
	TeamHITACHI.1	0.125	0.068	0.760	0.165	0.092	0.814
	TeamHITACHI.2	0.112	0.061	0.653	0.152	0.085	0.729
	HPI.1	0.106	0.059	0.520	0.151	0.086	0.620
Uncertainty Indicator	TeamHITACHI.2	0.561	0.496	0.647	0.672	0.612	0.746
	TeamHITACHI.1	0.514	0.693	0.408	0.655	0.802	0.553
	TeamHCMUS.1	0.252	0.169	0.494	0.386	0.275	0.646
	HPI.1	0.166	0.106	0.376	0.306	0.209	0.572
Course Class	TeamHITACHI.1	0.645	0.607	0.689	0.670	0.632	0.712
	TeamHITACHI.2	0.642	0.606	0.682	0.667	0.632	0.705
	TeamHCMUS.1	0.413	0.316	0.594	0.447	0.348	0.628
	HPI.1	0.226	0.153	0.435	0.283	0.196	0.510
Severity Class	TeamHITACHI.2	0.847	0.854	0.839	0.850	0.857	0.843
	TeamHITACHI.1	0.843	0.845	0.841	0.847	0.848	0.845
	TeamHCMUS.1	0.703	0.665	0.746	0.710	0.672	0.752
	HPI.1	0.364	0.306	0.448	0.396	0.336	0.483
Conditional Class	TeamHITACHI.1	0.638	0.744	0.559	0.801	0.869	0.743
	TeamHITACHI.2	0.548	0.478	0.643	0.729	0.669	0.800
	TeamHCMUS.1	0.307	0.225	0.484	0.441	0.340	0.625
	HPI.1	0.100	0.059	0.315	0.317	0.209	0.658
Generic Class	TeamHITACHI.1	0.225	0.239	0.213	0.304	0.320	0.289
	TeamHITACHI.2	0.192	0.385	0.128	0.263	0.484	0.181
	HPI.1	0.100	0.058	0.380	0.139	0.081	0.470
	TeamHCMUS.1	0.000	0.000	0.000	0.000	0.000	0.000
Body Location	TeamHITACHI.2	0.854	0.880	0.829	0.874	0.897	0.853
	TeamHITACHI.1	0.847	0.866	0.829	0.868	0.885	0.852
	TeamHCMUS.1	0.627	0.568	0.700	0.750	0.701	0.807
	HPI.1	0.134	0.298	0.086	0.363	0.611	0.258
Temporal Expression	TeamHCMUS.1	0.287	0.313	0.265	0.354	0.383	0.329
	TeamHITACHI.2	0.275	0.226	0.354	0.370	0.310	0.458
	TeamHITACHI.1	0.269	0.217	0.356	0.364	0.300	0.461
	HPI.1	0.000	0.000	0.000	0.000	0.000	0.000

Table 7. Evaluation in Task 3 (a) – part 1; baseline results are also provided. The best P@10 value for each team is emphasised.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret
baseline.bm25	0.6080	0.5680	0.6023	0.5778	0.3410	2346
baseline.dir	0.7240	*0.6800*	0.6926	0.6790	0.3789	2427
baseline.jm	0.4400	0.4480	0.4417	0.4510	0.2832	2399
baseline.tfidf	0.604	0.5760	0.5733	0.5641	0.3137	2326
COMPL_EN_Run.1	0.5184	0.4776	0.4896	0.4688	0.1775	1665
COMPL_EN_Run.5	0.5640	*0.5540*	0.5601	0.5471	0.2076	1828
CUNI_EN_RUN.1	0.5240	0.5060	0.5353	0.5189	0.3064	2562
CUNI_EN_RUN.5	0.5320	*0.5360*	0.5449	0.5408	0.3134	2556
CUNI_EN_RUN.6	0.5080	0.5320	0.5310	0.5395	0.2100	1832
CUNI_EN_RUN.7	0.5120	0.4660	0.5333	0.4878	0.1845	1676
DEMIR_EN_Run.1	0.6720	0.6300	0.6536	0.6321	0.3644	2479
DEMIR_EN_Run.5	0.7080	0.6700	0.6960	0.6719	0.3714	2493
DEMIR_EN_Run.6	0.6840	*0.6740*	0.6557	0.6518	0.3049	2281
DEMIR_EN_Run.7	0.6880	0.6120	0.6674	0.6211	0.3261	2404
ERIAS_EN_Run.1	0.5040	0.5080	0.4955	0.5023	0.3111	2537
ERIAS_EN_Run.5	0.5440	0.5280	0.547	0.5376	0.2217	2061
ERIAS_EN_Run.6	0.5720	*0.5460*	0.5702	0.5574	0.2315	2148
ERIAS_EN_Run.7	0.5960	0.5320	0.5905	0.5556	0.2333	2033
GRIUM_EN_Run.1	0.7240	0.7180	0.7009	0.7033	0.3945	2537
GRIUM_EN_Run.5	0.7680	*0.7560*	0.7423	0.7445	0.4016	2550
GRIUM_EN_Run.6	0.7480	0.7120	0.7163	0.7077	0.4007	2549
GRIUM_EN_Run.7	0.6920	0.6540	0.6772	0.6577	0.3495	2398
IRLabDAIICT_EN_Run.1	0.7120	*0.7060*	0.6926	0.6869	0.4096	2503
IRLabDAIICT_EN_Run.2	0.7040	0.7020	0.6862	0.6889	0.4146	2558
IRLabDAIICT_EN_Run.3	0.5480	0.5640	0.5582	0.5658	0.2507	2032
IRLabDAIICT_EN_Run.5	0.6680	0.6540	0.6523	0.6363	0.3026	2250
IRLabDAIICT_EN_Run.6	0.7320	0.6880	0.7174	0.6875	0.3686	2529
IRLabDAIICT_EN_Run.7	0.3160	0.2940	0.3110	0.2943	0.1736	1837
KISTI_EN_Run.1	0.7400	0.7300	0.7195	0.7235	0.3978	2567
KISTI_EN_Run.2	0.7320	*0.7400*	0.7191	0.7301	0.3989	2567
KISTI_EN_Run.3	0.7240	0.7160	0.7187	0.7171	0.3959	2567
KISTI_EN_Run.4	0.7560	0.7380	0.7390	0.7333	0.3971	2567
KISTI_EN_Run.5	0.7440	0.7280	0.7194	0.7211	0.3977	2567
KISTI_EN_Run.6	0.74400	0.7240	0.7218	0.7187	0.3971	2567
KISTI_EN_Run.7	0.7480	0.7260	0.7271	0.7233	0.3949	2567
miracl_en_run.1	0.6080	*0.5460*	0.6018	0.5625	0.1677	1189

Table 8. Evaluation in Task 3 (a) – part 2; baseline results are also provided. The best P@10 team value for each team is emphasised.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret
NIJM_EN_Run.1	0.5400	0.5740	0.5572	0.5708	0.3036	2330
NIJM_EN_Run.2	0.6240	*0.6180*	0.6188	0.6149	0.2825	2190
NIJM_EN_Run.3	0.5760	0.5960	0.5594	0.5772	0.2606	2154
NIJM_EN_Run.4	0.5760	0.5960	0.5594	0.5772	0.2606	2154
NIJM_EN_Run.5	0.5760	0.5880	0.5657	0.5773	0.2609	2165
NIJM_EN_Run.6	0.5120	0.5220	0.5332	0.5302	0.2180	1939
NIJM_EN_Run.7	0.5120	0.5220	0.5332	0.5302	0.2180	1939
RePaLi_EN_Run.1	0.6980	0.6612	0.6691	0.652	0.4054	2564
RePaLi_EN_Run.5	0.6920	*0.6740*	0.6927	0.6793	0.4021	2618
RePaLi_EN_Run.6	0.6880	0.6600	0.6749	0.6590	0.3564	2424
RePaLi_EN_Run.7	0.6720	0.6320	0.6615	0.6400	0.3453	2422
SNUMEDINFO_EN_Run.1	0.7720	0.7380	0.7337	0.7238	0.3703	2305
SNUMEDINFO_EN_Run.2	0.7840	*0.7540*	0.7502	0.7406	0.3753	2307
SNUMEDINFO_EN_Run.3	0.7320	0.6940	0.7166	0.6896	0.3671	2351
SNUMEDINFO_EN_Run.4	0.6880	0.6920	0.6562	0.6679	0.3514	2302
SNUMEDINFO_EN_Run.5	0.8160	0.7520	0.7749	0.7426	0.3814	2305
SNUMEDINFO_EN_Run.6	0.7840	0.7420	0.7417	0.7223	0.3655	2305
SNUMEDINFO_EN_Run.7	0.7920	0.7420	0.7505	0.7264	0.3716	2305
UHU_EN_Run.1	0.5760	0.5620	0.5602	0.5530	0.2624	2138
UHU_EN_Run.5	0.6040	*0.5860*	0.6169	0.5985	0.3152	2465
UHU_EN_Run.6	0.4880	0.5140	0.4997	0.5163	0.2588	2364
UHU_EN_Run.7	0.5560	0.5100	0.5378	0.5158	0.3009	2432
UIOWA_EN_Run.1	0.6880	*0.6900*	0.6705	0.6784	0.3589	2359
UIOWA_EN_Run.5	0.6840	0.6600	0.6579	0.6509	0.3226	2385
UIOWA_EN_Run.6	0.6760	0.6820	0.6380	0.6520	0.3259	2280
UIOWA_EN_Run.7	0.7000	0.6760	0.6777	0.6716	0.3452	2435
YORKU_EN_Run.1	0.4640	0.4360	0.4470	0.4305	0.1725	2296
YORKU_EN_Run.5	0.5840	*0.6040*	0.5925	0.5999	0.3207	2549
YORKU_EN_Run.6	0.0640	0.0600	0.0566	0.0560	0.0625	2531
YORKU_EN_Run.7	0.0480	0.0680	0.0417	0.0578	0.0548	2194

Table 9. Evaluation in Task 3 (b). Results for the cross lingual submissions are reported along with the corresponding English results. The best P@10 for each team-language is emphasised.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret
CUNI_EN_RUN.1	0.5240	0.5060	0.5353	0.5189	0.3064	2562
CUNI_EN_RUN.5	0.5320	*0.5360*	0.5449	0.5408	0.3134	2556
CUNI_EN_RUN.6	0.5080	0.5320	0.5310	0.5395	0.2100	1832
CUNI_EN_RUN.7	0.5120	0.4660	0.5333	0.4878	0.1845	1676
CUNI_CS_RUN.1	0.4400	0.4340	0.4361	0.4335	0.2151	1965
CUNI_CS_RUN.5	0.4920	*0.4880*	0.4830	0.4810	0.2399	2112
CUNI_CS_RUN.6	0.4680	0.4560	0.4928	0.4746	0.1573	1591
CUNI_CS_RUN.7	0.3360	0.3020	0.3534	0.3213	0.1095	1186
CUNI_DE_RUN.1	0.3837	0.400	0.3561	0.3681	0.1872	1806
CUNI_DE_RUN.5	0.4160	*0.4280*	0.3963	0.4058	0.2014	1935
CUNI_DE_RUN.6	0.3880	0.3820	0.4125	0.4024	0.1348	1517
CUNI_DE_RUN.7	0.3520	0.3200	0.3590	0.3330	0.1308	1556
CUNI_FR_RUN.1	0.4640	0.4720	0.4611	0.4675	0.2344	2056
CUNI_FR_RUN.5	0.4840	*0.4840*	0.4766	0.4776	0.2398	2064
CUNI_FR_RUN.6	0.4600	0.4560	0.4772	0.4699	0.1703	1531
CUNI_FR_RUN.7	0.3520	0.3240	0.3759	0.3520	0.1300	1313
SNUMEDINFO_EN_Run.1	0.7720	0.7380	0.7337	0.7238	0.3703	2305
SNUMEDINFO_EN_Run.5	0.8160	*0.7520*	0.7749	0.7426	0.3814	2305
SNUMEDINFO_EN_Run.6	0.7840	0.7420	0.7417	0.7223	0.3655	2305
SNUMEDINFO_EN_Run.7	0.7920	0.7420	0.7505	0.7264	0.3716	2305
SNUMEDINFO_CZ_Run.1	0.7837	0.7367	0.7128	0.6940	0.3473	2147
SNUMEDINFO_CZ_Run.5	0.7592	*0.7551*	0.6998	0.7011	0.3494	2147
SNUMEDINFO_CZ_Run.6	0.7388	0.7469	0.6834	0.6871	0.3395	2147
SNUMEDINFO_CZ_Run.7	0.7510	0.7367	0.6949	0.6891	0.3447	2147
SNUMEDINFO_DE_Run.1	0.7673	*0.7388*	0.6986	0.6874	0.3184	2087
SNUMEDINFO_DE_Run.5	0.7388	0.7347	0.6839	0.6790	0.3222	2087
SNUMEDINFO_DE_Run.6	0.7429	0.7286	0.6825	0.6716	0.3144	2087
SNUMEDINFO_DE_Run.7	0.7388	0.7122	0.6866	0.6645	0.3184	2087
SNUMEDINFO_FR_Run.1	0.7673	0.7429	0.7168	0.7077	0.3412	2175
SNUMEDINFO_FR_Run.5	0.7633	*0.7469*	0.7242	0.7090	0.344	2175
SNUMEDINFO_FR_Run.6	0.7592	0.7306	0.7121	0.6940	0.3320	2175
SNUMEDINFO_FR_Run.7	0.7469	0.7327	0.7078	0.6956	0.3363	2175

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., et al.: 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. 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)
6. 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)
7. 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)
8. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: MIE village of the future. (2012)
9. Koopman, B., Zuccon, G.: Relevation! an open source system for information retrieval relevance assessment. arXiv preprint (2013)
10. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
11. Robertson, S.E., Jones, S.: Simple, proven approaches to text retrieval. Technical Report 356, University of Cambridge (1994)
12. Smucker, M., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM '07. (2007) 623–632
13. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4) (2002) 422–446