ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval

Lorraine Goeuriot[1], Liadh Kelly[1], Wei Li[1], Joao Palotti[2], Pavel Pecina[3] Guido Zuccon[4] Allan Hanbury[2], Gareth J.F. Jones[1], and Henning Müller[5] *

[1] Dublin City University, Ireland, Firstname.Lastname@computing.dcu.ie
[2] Vienna University of Technology, Austria, palotti,hanbury@ifs.tuwien.ac.at
[3] Charles University in Prague, Czech Republic pecina@ufal.mff.cuni.cz
[4] Queensland University of Technology, Australia g.zuccon@qut.edu.au
[5] HES–SO, Sierre, Switzerland, henning.mueller@hevs.ch

Abstract. This paper presents the results of task 3 of the ShARe/CLEF eHealth Evaluation Lab 2014. This evaluation lab focuses on improving access to medical information on the web. The task objective was to investigate the effect of using additional information such as a related discharge summary and external resources such as medical ontologies on the effectiveness of information retrieval systems, in a monolingual (Task 3a) and in a multilingual (Task 3b) context. The participants were allowed to submit up to seven runs for each language (English, Czech, French, German), one mandatory run using no additional information or external resources, and three each using or not using discharge summaries.

Key words: Information retrieval, Evaluation, Medical information retrieval

1 Introduction

The goal of the ShARe/CLEF (Cross-Language Evaluation Forum) eHealth Evaluation Lab is to evaluate systems that support laypeople in searching for and understanding their health information [1]. It comprises three tasks. The specific use case considered is as follows: upon leaving the hospital, a patient receives a discharge summary. This describes the diagnosis and the treatment that they received in the hospital. Task 1 focuses 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 explores information extraction from clinical reports. Finally, this year's Task 3 further extends the 2013 information retrieval task, by cleaning the 2013 document collection and introducing a

* In alphabetical order, LG, LK led Task 3; WL, JP, PP & GZ contributed to the creation of the datasets, evaluation result generation and participant support activities; AH, GJFJ & HM were on the Task 3 organizing committee.

new query generation method and multilingual topics. This year then, Task 3 is split into Task 3a and Task 3b. Task 3a, similar to last year's Task 3, is a monolingual English retrieval task. Task 3b, adds a cross-lingual retrieval challenge to the lab, where participants must first translate parallel German, French and Czech queries into English before performing retrieval. The overall goal of Task 3 is to provide valuable and relevant documents to patients, so as to satisfy their health-related information needs. To evaluate systems that tackle this third task, we provide potential patient queries and a document collection containing various health and biomedical documents for task participants to create their search system. As is common in evaluation of information retrieval (IR), the test collection consists of documents, topics[6], and corresponding relevance judgements.

Searching for health advice is a common and important task performed by individuals on the web. Nearly 70% of search engine users in the US have conducted a web search for information about a specific disease or health problem [2]. While health IR is often considered as a domain-specific task, it is performed by a large variety of users, including various healthcare workers, but also, and increasingly commonly, by laypeople (e.g., patients and their relatives). This variety of potential information seekers, each characterized by different health knowledge, implies a broad range of information needs, and consequently a requirement for retrieval systems able to satisfy the health information needs of different categories of users.

The growing importance of health IR has provided the motivation for a number of evaluation campaigns focusing on health information. For example, the TREC (Text REtrieval Conference) Medical Records Track aims at identifying patient cohorts from medical reports to recruit for clinical trials [3]. In this task, topics include a particular disease/condition set and a particular treatment/intervention set; demographics or other characteristics may also be part of the topics (e.g., age group and hospitalization status). Moreover, the ImageCLEFmed tracks of the CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum) have created resources for the evaluation of image search in online resources or biomedical journal articles [4, 5]. However, while addressing different information needs (e.g., finding similar clinical cases vs. journal papers), these previous campaigns have targeted specific groups of users with expert health knowledge (e.g., clinicians and health researchers). The ShARe/CLEF eHealth Task 3 resembles other ad-hoc information retrieval tasks but with a focus on the information needs of laypeople and the types of queries they pose to express these needs. Results from the 2013 task [6] showed that this was a challenging task, with space for improvement and innovative techniques. Results from this year show considerable improvement over last year's results, both for the team submissions and the baseline, albeit on a new query set.

[6] A *topic* is considered to be an enriched version of a *query*, but both terms are used to refer to a topic in the paper.

The rest of this paper is organized as follows: Section 2 outlines the main IR evaluation campaigns on health topics. Section 3 describes the creation of the CLEF eHealth dataset, that is, the document collection, query generation, and relevance assessment. Section 4 presents the result sets and their evaluation and Section 5 the approaches used by task participants. Finally Section 6 concludes the paper.

2 Related Work

Previous research has considered the information needs of individuals seeking health advice on the web, but these studies mainly analyzed query logs from large commercial search engines [7]. To the best of our knowledge, no evaluation campaign has considered the information needs that patients may have regarding their health conditions and provided resources for evaluating IR systems for this task. Such lack of attention to this task arises, at least partially, due to the complexity of assessing the information needs: laypeople that search for health information on the web have very varied profiles, and their queries and searching time tend to be much shorter than those considered in past health IR benchmarks [8, 9].

OHSUMED, published in 1994, was the first collection containing medical data used for IR evaluation [10]. The collection contained around 350,000 abstracts from medical journals on the MEDLINE database over a period of five years (1987–1991) and two sets of topics: 63 topics manually generated and around 5,000 topics based on the controlled vocabulary thesaurus of the Medical Subject Headings[7] (concept name and definition). The collection was created for the TREC 2000 Filtering Track but also used for other research on health IR [11, 12].

The TREC Medical Records Track ran in 2011 and 2012 [3]. It was based on a collection of de-identified medical records (93,551 medical reports mapped into 17,264 visits) and queries (35 queries in 2011 and 50 in 2012) that resembled eligibility criteria of clinical studies. Records were grouped into visits, corresponding to a patient admission in the hospital; visits ranged in length from a few hours to in excess of a year. The goal of the track was to find patient cohorts that are relevant to the criteria for recruitment as populations in comparative effectiveness studies. In 2014, TREC organized a new medical evaluation challenge, called TREC Clinical Decision Support Track[8]. The focus of the track is the retrieval of biomedical articles relevant for answering generic clinical questions about medical records. Participants are provided with short case reports, as idealized representations of actual medical records. They have to retrieve biomedical articles that answer questions related to several types of clinical information needs based on the report.

In 2013, CLEF hosted a workshop and challenge focusing on multilingual biomedical named entities recognition, CLEF-ER[13]. Their challenge was based

[7] http://www.ncbi.nlm.nih.gov/mesh/

[8] http://www.trec-cds.org/

on a parallel corpus in English, French, German, Spanish, and Dutch, composed of patent texts, titles of Medline abstracts and EMEA documents. The goal of the task was to identify concepts by their CUIs (Concept Unique Identifiers) in the documents, using biomedical terminological resources, and an annotated English corpus.

3 Task 3 Description

The data set provided to participants comprises a document collection of around one million documents (web pages from medical web sites), 50 parallel topics (in English (EN), Czech (CS), French (FR), and German (DE)), which were developed by medical experts in English and translated into CS, FR and DE, and the corresponding relevance information. In addition to TREC-style title and description fields, the topics contain an additional field discharge-summary, which contains the discharge report which the patient's query stemmed from.

The data was provided to participants after signing an agreement, through the PhysioNet website. As test data, five parallel training topics (in EN, CS, FR, and DE) together with corresponding relevance assessment were released.

In this section we describe each part of the task dataset.

3.1 Document Collection

A large web crawl of health resources is used as the corpus for this task. This is an updated version of the web crawl released for CLEFeHealth Task 3 2013. In this updated version further efforts have been made to clean the document collection, by removing duplicate documents with the same URL and fixing detected errors in HTML.

The crawl contains about one million documents, which have been made available to CLEF eHealth through the Khresmoi project [15]. This collection consists of web pages covering a broad range of health topics, targeted at both the general public and healthcare professionals. These domains consist predominantly of health and medicine websites that have been certified by the Health on the Net (HON) Foundation[9] as adhering to the HONcode principles[10] (approximately 60–70% of the collection), as well as other commonly used health and medicine websites such as Drugbank[11], Diagnosia[12] and Trip Answers[13]. The crawled documents are provided in the dataset in their raw HTML (Hyper Text Markup Language) format along with their uniform resource locators (URL). The dataset is made available for download on the web to registered participants on a secure password-protected server.

9 http://www.healthonnet.org
10 http://www.hon.ch/HONcode/Patients-Conduct.html
11 http://www.drugbank.ca/
12 http://www.diagnosia.com/
13 http://www.tripanswers.org/

3.2 Discharge Summaries

Novel methods to generate contextualized statements of patient information needs were used. These are based on realistic short query statements created in the context of patient discharge summaries. The discharge summaries can be considered as a description of the context in which the patient has been diagnosed with a given disorder and has written a query. The discharge summaries originate from the de-identified MIMIC-II database[14] (Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5). They are, together with annotations, CLEF eHealth task 2 dataset [16].

Discharge summaries are semi-structured reports with the following appearance:

```
Admission Date:  [**2014-03-28**]
Discharge Date:   [**2014-04-08**]
Date of Birth:  [**1930-09-21**]
Sex:   F
Service: CARDIOTHORACIC
Allergies:
 Patient recorded as having No Known Allergies to Drugs

Attending:[**Attending Info 565**]
Chief Complaint: Chest pain
Major Surgical or Invasive Procedure:
 Coronary artery bypass graft 4.
History of Present Illness:
 83 year-old woman, patient of Dr. [**First Name4
 (NamePattern1) **] [**Last Name (NamePattern1) 5005**],
 Dr. [**First Name (STitle) 5804**] [**Name (STitle)
 2275**], with increased SOB with activity, left shoulder
 blade/back pain at rest, + MIBI, referred for cardiac
 cath. This pleasant 83 year-old patient notes becoming
 SOB when walking up hills or inclines about one year
 ago. This SOB has progressively worsened and she is now
 SOB when walking [**01-19**] city block (flat surface).
[...]

Past Medical History:
 arthritis; carpal tunnel; shingles right arm 2000;
 needs right knee replacement; left knee replacement
 in [**2010**]; thyroidectomy 1978; cholecystectomy
 [**1981**]; hysterectomy 2001; h/o LGIB 2000-2001
 after taking baby ASA; 81 QOD
[...]
```

3.3 Topics

In this section we describe the creation of the initial English topic set used in Task 3a, and the translation of this topic set into Czech, French and German to form a parallel topic corpus for use in Task 3b.

English Topics The queries used in the task aim to model those used by laypeople (i.e., patients, their relatives or other representatives) to find out more about their disorders, once they have examined a discharge summary.

[14] http://mimic.physionet.org

Topics to be used in this task have been created by experts (each expert was a registered nurse and clinical documentation researcher) involved in the CLEF eHealth consortium. This solution has been chosen in place of recruiting patients because of the issues involved with recruitment and privacy. We believe that, being on a daily basis in contact with patients receiving treatments and discharge summaries, nurses are familiar with patients' information needs and patient profiles.

Topics have been manually created by the experts given discharge summaries, and the discharge diagnosis. Last year's queries were generated from randomly selected disorders. Therefore, the disorder was often not central enough in the discharge summary for it to provide useful IR contextual information [6]. This year, queries were built based on one of the main disorders, identified from the discharge summary, which the patient was hospitalized for. Discharge summaries are semi-structured documents, and the discharge diagnosis is a field that can be found in 85% of the discharge summaries. The discharge diagnosis contains on average 3 disorders. From these three, the experts selected one which a patient may have questions on. For discharge summaries which had no discharge diagnosis, experts selected a main disorder within the discharge summary, which a patient may have questions on. Using the pairs of disorder and associated discharge summary, the experts developed a set of patient queries (and criteria for judging the relevance of documents to the queries, for use in the relevance assessment task described in the next section). Queries are provided in a standard TREC format, consisting of a topic title (text of the query), description (longer description of what the query means), a narrative (expected content of the relevant documents), and a profile (brief description of the patient).

The following example outlines a query:

```
<query>
  <title> thrombocytopenia treatment corticosteroids
    length </title>
  <desc> How long should be the corticosteroids treatment
    to cure thrombocytopenia? </desc>
  <narr> Documents should contain information about
    treatments of thrombocytopenia, and especially
    corticosteroids. It should describe the treatment,
    its duration and how the disease is cured using it.
    <scenario> The patient has a short-term disease, or
     has been hospitalised after an accident (little to
     no knowledge of the disorder, short-term treatment)
    </scenario>
    <profile> Professional female </profile>
  </narr>
</query>
```

With this approach, five training and fifty test queries have been generated for use in the task.

Translated Topics For the purpose of Task 3b, the original topics in English were manually translated into Czech, German, and French. Based on our previous experience with manual translation of medical user queries [17, 18] the translation was performed in three phases: First, the topics were translated from English to the target languages by medical experts (one translator per language, not necessarily native speakers but fluent in the target languages). Second, the translations were reviewed by language experts (native speakers or people with a university degree in that language) and any language-related issues (typos, grammar, etc.) were resolved. Third, any terminology issues were consulted with the original translators and resolved together with the language experts.

We asked the translators (and reviewers) to produce translations while grammatically correct, preserve meaning and use terminology adequate to the technical level of the original topic descriptions. Unlike the original topics, the resulting translations do not contain any grammatical errors and typos.

3.4 Relevance Assessment

For this year's task, relevance judgements were collected from professional assessors (but not medical experts). We used Relevation! [19][15] to manage the collection of relevance assessments for documents in the assessment pool, where each document was judged by one assessor.

To form the assessment pool, we selected the top ten documents obtained from the participants' baseline runs (run 1), their top-two priority runs using discharge summaries (runs 2 and 3), and their top-two priority runs not using discharge summaries (runs 5 and 6). This resulted in a pool of 6,800 documents, in line with the size of the pool for the 2013 task. The relevance assessment was based on a four point scale. The relevance grades are: (0) irrelevant, (1) on topic but unreliable, (2) relevant, (3) highly relevant. These relevance grades are mapped into a binary scale, with grades 0 and 1 corresponding to the binary grade 0 (irrelevant) and grades 2 and 3 corresponding to the binary grade 1 (relevant). The graded relevance assessment yielded 0: 3,044, 1: 547, 2: 974, 3: 2,235 documents. The binary relevance assessments yielded 0: 3,591 non-relevant and 1: 3,209 relevant documents. This year's assessment exercise yielded more relevant documents per topic than last year: 64.18 relevant documents per topic on average compared to last year's 37.56.

Relevance assessments for the five training queries were formed based on pooled sets generated using the Vector Space Model [20] and Okapi BM25 [21]. Assessments for these five training queries were conducted by two Finnish nurses. Each document was assessed by one person. Training queries were distributed to participants before the test queries were released.

4 Results

For this task, the participants were allowed to submit up to seven runs for the English monolingual retrieval task, Task 3a. These runs comprised, one mandatory run using no additional information or external resources (run 1), three runs using the discharge summary and any other external resource (runs 2-4), and three using external resources but not using the discharge summaries (run

[15] http://ielab.github.io/relevation/

5-7). Among each set of additional runs, one had to use only the title and the description fields of the query. Participants were also asked to rank their runs 2-4 and 5-7 according to their importance. For the cross-language information retrieval task, Task 3b, participants could submit up to seven runs for each language (Czech-English, French-English, German-English). These runs had the same make-up as those in Task 3a.

4.1 Participants

This year, 91 groups registered for the task, 25 obtained access to the data and 14 submitted run(s) for task 3. The groups are from 11 countries in 4 continents as listed in Table 1. While only one group from Europe participated last year, this year the European groups formed the majority.

Table 1. Participants for task 3a and 3b and their total number of submissions.

Continent	Country	Team Name	Runs Submitted	
			Task 3a	Task 3b
Africa	Tunisia	Miracl	1	-
America	Canada	GRIUM	4	-
	Canada	YORKU	4	-
	USA	UIOWA	4	-
Asia	India	IRLabDAIICT	6	-
	South Korea	SNUMEDINFO	7	4 runs/language
	South Korea	KISTI	7	-
Europe	Thailand	CSKU/COMPL	2	-
	Czech Republic	CUNI	4	4 runs/language
	France	ERIAS	4	-
	France	RePaLi	4	-
	Netherlands	Nijmegen	7	-
	Spain	UHU	4	-
	Turkey	DEMIR	4	-

Teams submitted in total 62 runs for task 3a in which 11 used discharge summaries (from teams IRLabDAIICT, SNUMEDINFO, KISTI and Nijmegen). For task 3b, 24 runs were submitted by two groups.

4.2 Evaluation Metrics

We examined all documents in runs 1, 2, 3, 5 and 6 from Tasks 3a and 3b up to rank 10 for relevance. The two major evaluation metrics are therefore metrics at a cut-off of up to 10 documents, i.e. P@5, P@10, NDCG@5, and NDCG@10. In addition, we considered MAP as an evaluation metric, but we are aware that MAP is unreliable because only the top ten documents have been assessed. Nevertheless, we wanted to report a measure covering the full set of up to 1000 retrieved documents. We also report the number of relevant and retrieved documents in the top 1000 results as a more recall-oriented measure.

Performance metrics are computed with the standard *trec_eval* tool[16] using the following commands:

– -c -M1000 qrels.clef2014.test.bin.txt runName
– -c -M1000 -m ndcg_cut qrels.clef2014.test.graded.txt runName

We are aware that the performance metrics for other runs might be unreliable compared to that of runs 1, 2, 3, 5 and 6. However, this situation is common for IR lab evaluations, where additional experiments on an existing data set typically do not include re-assessment of documents previously not retrieved or relevance assessment of additional documents.

4.3 Baseline System

For comparison, we created our own baseline experiments by implementing a number of information retrieval baselines: tf.idf (baseline.tfidf), BM25 (baseline.bm25), language modeling with Jelinek-Mercer smoothing (baseline.jm), and language modeling with Dirichlet smoothing (baseline.dir). These methods do not incorporate any domain-specific adaptations. We used the implementations of the above methods made available in the Indri toolkit[17]. Indri was also used to parse the HMTL documents and for stemming (with Krovetz stemming, also applied to queries). A stop list was applied to the queries but not to the documents.

4.4 Evaluation Results

The official results for all runs submitted to Task 3 (both 3a and 3b) and for our baseline experiments (highlighted in italics) are shown in Tables 2 and 2, ordered by decreasing P@10 (Task 3's primary measure). Comparing the participants' results with respect to P@10 we observe that, for each team, the best effectiveness is often achieved when no discharge summaries are considered (runs 5, 6, 7 and 1, which is the teams' baseline); teams KISTI and NIJM are an exception to this trend. A similar result was found also in the 2013 campaign, with most of the teams achieving the highest effectiveness when not using discharge summaries.

Two teams submitted to the cross-lingual Task 3b: SNUMEDINFO and CUNI. The results obtained by the SNUMEDINFO team when using the cross-lingual queries demonstrate comparable results to the corresponding submissions when using English queries: in some cases cross-lingual queries yield even higher results than the original English queries (e.g. `SNUMEDINFO_CZ_Run.5` vs. `SNUMEDINFO_EN_Run.5`), and these are comparable to the best results obtained for the original English queries (Task 3a). This is not the case though for team

[16] http://trec.nist.gov/trec_eval/
[17] www.lemurproject.org

CUNI, whose cross-lingual submissions generally yield less effectiveness than the corresponding Task 3a submissions.

The best result in last year's task was obtained by TeamMayo, with a P@10 of 0.5180. This year's best run is obtained by team SNUMEDINFO with a P@10 of 0.7560. Even the baselines have considerably improved on 2014 dataset. Several changes have been made between the two tasks: the document collection has been reduced, and the query generation strategy has changed (from a randomly selected disorder to the main one). One hypothesis to explain the increase could be the fact that the topics are simpler, in the way that they correspond to main disorders, that are potentially more frequent and more searched in general. Further analysis is required to explain this improvement.

5 Approaches Used

In this section we describe the approaches used by each team, and summarize findings from their analysis. Table 4 provides a condensed view of the techniques and resources used by each team.

Team CSKU-COMPL [22] used the vector space retrieval model of Lucene as baseline. As improvement, they proposed a simple pseudo-relevance feedback method which used the Genomic collection as external resource to perform query expansion. The expansion terms selection is based on the Rocchio's formula with dynamic tunable parameter of Pseudo-relevance feedback. Their best run obtained P@10 of 0.5540.

Team CUNI [23] participated in both tasks 3a and 3b, using only the query titles and the Terrier platform (Hiemstra retrieval model) as their baseline. They employed various methods for data cleaning and the simplest one, removing only the HTML tags, had the best results. Their best run for task 3a used suggestions from the MedlinePlus dictionary to fix typos in the queries (P@10 of 0.5360). They also employed query expansion adding the top ten highest terms from the top 3 ranked documents, but this did not improve the results. For task 3b, only one step was included, which was the translation of query titles using Khresmoi translator system. Their best run here obtained P@10 of 0.4880 for Czech.

Team DEMIR [24] has as baseline the Terrier system. For each query they predict whether query expansion is likely to improve retrieval performance or not. Prediction is performed using a Naive Bayes classifier trained on the CLEF eHealth 2013 test collection and features extracted from the queries and statistics obtained from the collection. Their best result achieved P@10 of 0.67.

Team ERIAS [25] used the Vector Space Model in Lucene, indexing both unigrams and bigrams for their baseline. The baseline system uses only the query title as the query and uses no external resources. Other runs include query expansion using synonymous terms and descendants from MeSH and the UMLS.

Table 2. Retrieval effectiveness of the top-45 runs submitted to Task 3 (both 3a and 3b). Runs are ordered by decreasing P@10. Baseline results are highlighted in italics and the best results for each evaluation measure marked in bold.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret
GRIUM_EN_Run.5	0.7680	**0.7560**	0.7423	**0.7445**	0.4016	2550
SNUMEDINFO_CZ_Run.5	0.7592	0.7551	0.6998	0.7011	0.3494	2147
SNUMEDINFO_EN_Run.2	0.7840	0.7540	0.7502	0.7406	0.3753	2307
SNUMEDINFO_EN_Run.5	**0.8160**	0.7520	**0.7749**	0.7426	0.3814	2305
SNUMEDINFO_CZ_Run.6	0.7388	0.7469	0.6834	0.6871	0.3395	2147
SNUMEDINFO_FR_Run.5	0.7633	0.7469	0.7242	0.7090	0.3440	2175
SNUMEDINFO_FR_Run.1	0.7673	0.7429	0.7168	0.7077	0.3412	2175
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
KISTI_EN_Run.2	0.7320	0.7400	0.7191	0.7301	0.3989	2567
SNUMEDINFO_DE_Run.1	0.7673	0.7388	0.6986	0.6874	0.3184	2087
KISTI_EN_Run.4	0.7560	0.7380	0.7390	0.7333	0.3971	2567
SNUMEDINFO_EN_Run.1	0.7720	0.7380	0.7337	0.7238	0.3703	2305
SNUMEDINFO_CZ_Run.1	0.7837	0.7367	0.7128	0.6940	0.3473	2147
SNUMEDINFO_CZ_Run.7	0.7510	0.7367	0.6949	0.6891	0.3447	2147
SNUMEDINFO_DE_Run.5	0.7388	0.7347	0.6839	0.6790	0.3222	2087
SNUMEDINFO_FR_Run.7	0.7469	0.7327	0.7078	0.6956	0.3363	2175
SNUMEDINFO_FR_Run.6	0.7592	0.7306	0.7121	0.6940	0.3320	2175
KISTI_EN_Run.1	0.7400	0.7300	0.7195	0.7235	0.3978	2567
SNUMEDINFO_DE_Run.6	0.7429	0.7286	0.6825	0.6716	0.3144	2087
KISTI_EN_Run.5	0.7440	0.7280	0.7194	0.7211	0.3977	2567
KISTI_EN_Run.7	0.7480	0.7260	0.7271	0.7233	0.3949	2567
KISTI_EN_Run.6	0.7440	0.7240	0.7218	0.7187	0.3971	2567
GRIUM_EN_Run.1	0.7240	0.7180	0.7009	0.7033	0.3945	2537
KISTI_EN_Run.3	0.7240	0.7160	0.7187	0.7171	0.3959	2567
SNUMEDINFO_DE_Run.7	0.7388	0.7122	0.6866	0.6645	0.3184	2087
GRIUM_EN_Run.6	0.7480	0.7120	0.7163	0.7077	0.4007	2549
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
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
UIOWA_EN_Run.1	0.6880	0.6900	0.6705	0.6784	0.3589	2359
IRLabDAIICT_EN_Run.6	0.7320	0.6880	0.7174	0.6875	0.3686	2529
UIOWA_EN_Run.6	0.6760	0.6820	0.6380	0.6520	0.3259	2280
baseline.dir	*0.7240*	*0.6800*	*0.6926*	*0.6790*	*0.3789*	*2427*
UIOWA_EN_Run.7	0.7000	0.6760	0.6777	0.6716	0.3452	2435
DEMIR_EN_Run.6	0.6840	0.6740	0.6557	0.6518	0.3049	2281
RePaLi_EN_Run.5	0.6920	0.6740	0.6927	0.6793	0.4021	**2618**
DEMIR_EN_Run.5	0.7080	0.6700	0.6960	0.6719	0.3714	2493
RePaLi_EN_Run.1	0.6980	0.6612	0.6691	0.6520	0.4054	2564
RePaLi_EN_Run.6	0.6880	0.6600	0.6749	0.6590	0.3564	2424
UIOWA_EN_Run.5	0.6840	0.6600	0.6579	0.6509	0.3226	2385
GRIUM_EN_Run.7	0.6920	0.6540	0.6772	0.6577	0.3495	2398
IRLabDAIICT_EN_Run.5	0.6680	0.6540	0.6523	0.6363	0.3026	2250
RePaLi_EN_Run.7	0.6720	0.6320	0.6615	0.6400	0.3453	2422

Table 3. Retrieval effectiveness of the bottom-45 runs submitted to Task 3 (both 3a and 3b). Runs are ordered by decreasing P@10. Baseline results are highlighted in italics and the best results for each evaluation measure marked in bold.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret
DEMIR_EN_Run.1	0.6720	0.6300	0.6536	0.6321	0.3644	2479
NIJM_EN_Run.2	0.6240	0.6180	0.6188	0.6149	0.2825	2190
DEMIR_EN_Run.7	0.6880	0.6120	0.6674	0.6211	0.3261	2404
YORKU_EN_Run.5	0.5840	0.6040	0.5925	0.5999	0.3207	2549
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
UHU_EN_Run.5	0.6040	0.5860	0.6169	0.5985	0.3152	2465
baseline.tfidf	*0.6040*	*0.5760*	*0.5733*	*0.5641*	*0.3137*	*2326*
NIJM_EN_Run.1	0.5400	0.5740	0.5572	0.5708	0.3036	2330
baseline.bm25	*0.6080*	*0.5680*	*0.6023*	*0.5778*	*0.3410*	*2346*
IRLabDAIICT_EN_Run.3	0.5480	0.5640	0.5582	0.5658	0.2507	2032
UHU_EN_Run.1	0.5760	0.5620	0.5602	0.553	0.2624	2138
COMPL_EN_Run.5	0.5640	0.5540	0.5601	0.5471	0.2076	1828
ERIAS_EN_Run.6	0.5720	0.5460	0.5702	0.5574	0.2315	2148
miracl_en_run.1	0.6080	0.5460	0.6018	0.5625	0.1677	1189
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
ERIAS_EN_Run.7	0.5960	0.5320	0.5905	0.5556	0.2333	2033
ERIAS_EN_Run.5	0.5440	0.5280	0.5470	0.5376	0.2217	2061
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
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
ERIAS_EN_Run.1	0.5040	0.5080	0.4955	0.5023	0.3111	2537
CUNI_EN_RUN.1	0.524	0.5060	0.5353	0.5189	0.3064	2562
CUNI_CS_RUN.5	0.4920	0.4880	0.4830	0.4810	0.2399	2112
CUNI_FR_RUN.5	0.4840	0.4840	0.4766	0.4776	0.2398	2064
COMPL_EN_Run.1	0.5184	0.4776	0.4896	0.4688	0.1775	1665
CUNI_FR_RUN.1	0.4640	0.4720	0.4611	0.4675	0.2344	2056
CUNI_EN_RUN.7	0.5120	0.4660	0.5333	0.4878	0.1845	1676
CUNI_CS_RUN.6	0.4680	0.4560	0.4928	0.4746	0.1573	1591
CUNI_FR_RUN.6	0.4600	0.4560	0.4772	0.4699	0.1703	1531
baseline.jm	*0.4400*	*0.4480*	*0.4417*	*0.4510*	*0.2832*	*2399*
YORKU_EN_Run.1	0.4640	0.4360	0.4470	0.4305	0.1725	2296
CUNI_CS_RUN.1	0.4400	0.4340	0.4361	0.4335	0.2151	1965
CUNI_DE_RUN.5	0.4160	0.4280	0.3963	0.4058	0.2014	1935
CUNI_DE_RUN.1	0.3837	0.400	0.3561	0.3681	0.1872	1806
CUNI_DE_RUN.6	0.3880	0.3820	0.4125	0.4024	0.1348	1517
CUNI_FR_RUN.7	0.3520	0.3240	0.3759	0.3520	0.1300	1313
CUNI_DE_RUN.7	0.3520	0.3200	0.3590	0.3330	0.1308	1556
CUNI_CS_RUN.7	0.3360	0.3020	0.3534	0.3213	0.1095	1186
IRLabDAIICT_EN_Run.7	0.3160	0.2940	0.3110	0.2943	0.1736	1837
YORKU_EN_Run.7	0.0480	0.0680	0.0417	0.0578	0.0548	2194
YORKU_EN_Run.6	0.0640	0.0600	0.0566	0.0560	0.0625	2531

Team	BaseSE	IR Model	DS	Query Expansion	External
CSKU	Lucene	VSM		PRF	Medline
CUNI	Terrier	Hiemstra		PRF system	Khresmoi MT
DEMIR	Terrier	VSM		KL expansion	Weka to classify queries
ERIAS	Lucene	VSM		Synonyms	MeSH, UMLS, Metamap
GRIUM	Indri	LM		Mutual Information	UMLS, Metamap
IRLabDAIICT	Indri	Vary	✓	Query-likelihood, Blind RF	Metamap, MeSH
KISTI	Lucene	LM	✓	Abbreviations and PRF	-
MIRACL	Terrier	VSM		-	-
nijmegen	Indri	LM	✓	Kullback-Leibler divergence, synonyms	UMLS
RePaLi	Indri	LM		synonyms, abbreviations	UMLS, FASTR, YATEA, Ogmios NLP
SNUMEDINFO	Indri	LM	✓	Intersection of preferred terms and DS	Metamap, UMLS
UHU	Lucene	?		synonyms, related terms	Metamap, MeSH, Tika
UIOWA	Indri	LM		MRF, PRF	GeniaSS
YORKU	Terrier	Vary		-	-

Table 4. Summarized view of the methods used by each team

For identifying medical terms in queries, a method has been developed that focuses on the most specific terms, i.e. only medical terms not sub-parts of other medical terms. Their best run obtained a P@10 of 0.5460.

Team GRUIM [26] experimented with the use of the UMLS Metathesarus to explore the effectiveness of concept-based retrieval techniques. Their baseline was based on Indri and Language Model with Dirichlet smoothing. They used Metamap to annotate the documents and extract the medical concept. They also experiment with query expansion using mutual information to determine related concepts. Their best run obtained a P@10 of 0.75.

Team IRLABDAIICT [27] indexed the document collection using Indri and used the query likelihood model as their baseline. Other runs compared the Okapi Model with the query likelihood model. They also experimented with using the discharge summaries combined with MeSH terminology for query expansion. Their best run was the baseline, which obtained P@10 of 0.70.

Team KISTI [28] proposed a multiple-stage re-ranking method. Their baseline used Lucene and query-likelihood with Dirichlet smoothing. It focuses on using various retrieval techniques rather then using external resources and NLP techniques. The sequential steps used are (i) query expansion with abbreviations, (ii) query expansion with the discharge summary, (iii) clustering-based document scoring, (iv) centrality-based document scoring using implicit links among documents, and (v) pseudo relevance feedback. Their best run obtained a P@10 of 0.74, which applied steps (i), (ii) and (v).

Team MIRACL [29] based their submissions on the Terrier retrieval system with fairly standard settings for tokenization, stop word removal and stemming. Their only run used a standard Vector Space Model, obtaining a MAP of 0.17 and a P@10 of 0.55.

Team Nijmegen [30] used the Language Modeling retrieval model of the Indri search engine with Pseudo-Relevance feedback as their baseline. They employed the Kullback-Leibler divergence for informativeness and phraseness method to expand the query with terms from the discharge summaries (runs 2 to 4) and UMLS-thesaurus (runs 5 to 7). The best result was found for run 4, where only the discharge summaries were used for query expansion (P@10 of 0.6540).

Team RePALI [31] also opted for the Indri system as a baseline (parameters estimated on the 2013 dataset), and experimented with various methods of incorporating morpho-syntactic variants, lexical inclusion and hierarchical relations, and abbreviations. However, results were inconsistent across the query set with the reasons for this not being clear. Their best run obtained a P@10 of 0.67.

Team SNUMEDINFO [32] submitted to both Tasks 3a and 3b. As baseline, they used the Indri retrieval system with Dirichlet smoothing language model. They experimented with query expansion using the Metamap system, in which candidate expansion keywords were filtered against the discharge summary associated with the original query. They also experimented with learning to rank based on random forests. They extracted features such as the "quality feature", which, by counting how many terms from a pre-compiled list appear in a document, attempts to estimate the reliability of the medical information presented in the document. Their best run for Task 3a obtained a P@10 of 0.75. Their cross-lingual submissions were based on the use of Google Translate, and their best run here obtained a P@10 of 0.75 for Czech.

Team UHU (LABERINTO) [33] used a standard system built on Lucene and experimented with methods for term boosting and query expansion. They submitted 4 runs not using the discharge summaries. In run 5, a boosting factor of 1.5 was applied to query terms which appear in UMLS, which increased P@10 from the baseline of 0.56 to 0.58. Query expansion, realized by adding MeSH descriptors for query terms appearing both in title and description, did not improve the baseline results.

Team UIOWA [34] included all webpage content in their document index, as opposed to just body text. They used Indri to generate their baseline. The other approaches they explored performed worse than this baseline (P@10 of 0.69). They experimented with pseudo relevance feedback and using the Markov Random Field Model with medical phrase bigrams extracted from MetaMap for query expansion.

Team YORKU [35] has as the core of their approach the use of Learning to Rank with a total of 231 features from multiple information retrieval models and different parameter settings. The group submitted several runs, in which they compare binary and graded relevance information, as well as the use of different machine learning algorithms. Their best run obtained a P@10 of 0.60.

6 Conclusions

In this second year of the ShARe/CLEF eHealth2014 evaluation lab Task 3, there was strong take-up in the community with 14 groups submitting runs to the task. The challenge of developing retrieval techniques for layperson medical queries proved difficult.

Overall, we observed a considerable improvement over 2013 results, both for the team runs and the baselines. The best run for task 3a was submitted by team GRIUM, with a P@10 of 0.7560 and a NDCG@10 of 0.7445. The best run for task 3b was submitted by team SNUMEDINFO on the Czech topics, with P@10 of 0.7551 and NDCG@10 of 0.7011 (their P@10 is slightly higher for Czech topics than for English ones). The three best teams use language modelling retrieval methods, perform some query expansion and two of them use UMLS. The best team for task 3b used Google Translate[18] to translate the queries.

This year, we implemented several state-of-the-art baselines. The highest performances are achieved using language models with Dirichlet smoothing.

Four teams submitted runs using the discharge summaries. Two of the top-10 runs (ranked with P@10) use them: SNUMEDINFO and KISTI. Moreover, all the runs using discharge summaries for these two teams obtain higher results than their runs without discharge summaries. This is an improvement over 2013, where no team managed to improve their results with the discharge summaries. Our new topic generation strategy proved to be more accurate, and discharge summaries seem to bring useful contextual information to better retrieve documents.

Given the success of the first two years of the task, we anticipate even more interest in next year's campaign. In the third year of this task, we will explore new topic generation strategies, based on our related research on automatic generation of queries [36] and analysis of query complexity [37]. Moreover, we intend to perform more analysis work to better understand the task results and IR methods to answer laypeople medical information needs.

Acknowledgement

Task 3 of the ShARe/CLEFeHealth2013 evaluation lab has been supported in part by the Khresmoi project, funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 257528. We acknowledge the time given to perform the relevance assessment task. We want to thank the following individuals: Riitta Danielsson-Ojala (University of Turku, Finland), Sanna Salanterä (University of Turku, Finland) for creating the queries and conducting relevance assessment, and Ondrej Dusek (Charles University), Brendan Hegarty (Dublin City University), Jaroslava Hlavacova (Charles University), John Hodmon (Dublin City University), Michal Novak (Charles University), David Racca (Dublin City University), Rudolf Rosa and Daniel Zeman (Charles University) for their help on the relevance assessment. We also acknowledge the time given by Margit Hanbury to check the German translations.

References

1. Kelly, L., Goeuriot, L., Suominen, H., Schrek, T., Leroy, G., Mowery, D.L., Velupillai, S., Chapman, W.W., Martinez, D., Zuccon, G., Palotti, J.: Overview of the share/clef ehealth evaluation lab 2014. In: Proceedings of CLEF 2014. Lecture Notes in Computer Science (LNCS), Springer (2014)
2. Fox, S.: Health topics: 80% of internet users look for health information online. Technical report, Pew Research Center (February 2011)
3. Voorhees, E.M., Tong, R.M.: Overview of the TREC 2011 medical records track. In: Proceedings of TREC, NIST (2011)

[18] http://translate.google.com

4. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A.G.S., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum). (2011)
5. Müller, H., Clough, P., Deselaers, T., Caputo, B., eds.: ImageCLEF — Experimental Evaluation in Visual Information Retrieval. Volume 32 of The Information Retrieval Series. Springer (2010)
6. Goeuriot, L., Jones, G.J.F., 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: CLEF online working notes. (2013)
7. White, R., Horvitz, E.: Cyberchondria: Studies of the escalation of medical concerns in web search. Technical report, Microsoft Research (2008)
8. Boyer, C., Gschwandtner, M., Hanbury, A., Kritz, M., Pletneva, N., Samwald, M., Vargas, A.: Use case definition including concrete data requirements (D8.2). public deliverable, Deliverable of the Khresmoi EU project (2012)
9. Suominen, H., ed.: The Proceedings of the CLEFeHealth2012 – the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis. NICTA (2012)
10. Hersh, W.R., Buckley, C., Leone, T.J., Hickam, D.H.: OHSUMED: An interactive retrieval evaluation and new large test collection for research. In: Proceedings of SIGIR '94. (1994) 192–201
11. Claveau, V.: Unsupervised and semi-supervised morphological analysis for information retrieval in the biomedical domain. In: Proceedings of COLING. (2012) 629–645
12. Koopman, B., Zuccon, G., Bruza, P., Sitbon, L., Lawley, M.: An evaluation of corpus-driven measures of medical concept similarity for information retrieval. In: Proceedings of CIKM 2012. (2012)
13. Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E.M., Bui, C., Hellrich, J., Lewin, I., Milward, D., Poprat, M., Jimeno-Yepes, A., Hahn, U., Kors, J.A.: Multilingual semantic resources and parallel corpora in the biomedical domain: the clef-er challenge. In: CLEF online working notes. (2013)
14. Goeuriot, L., Kelly, L., Jones, G., Zuccon, G., Suominen, H., Hanbury, A., Müller, H., Leveling, J.: Creation of a new evaluation benchmark for information retrieval targeting patient information needs. In Song, R., Webber, W., Kando, N., Kishida, K., eds.: Proceedings of the 5th International Workshop on Evaluating Information Access (EVIA), a Satellite Workshop of the NTCIR-10 Conference, Tokyo/Fukuoka, Japan, National Institute of Informatics/Kijima Printing (2013)
15. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: MIE village of the future. (2012)
16. Mowery, D.L., Velupillai, S., South, B.R., Christensen, L., Martinez, D., Kelly, L., Goeuriot, L., Elhadad, N., Pradhan, S., Savova, G., Chapman, W.W.: Task 2: Share/clef ehealth evaluation lab 2014. In: Proceedings of CLEF 2014. (2014)
17. Urešová, Z., Hajič, J., Pecina, P., Dušek, O.: Multilingual test sets for machine translation of search queries for cross-lingual information retrieval in the medical domain. In Chair), N.C.C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (2014)
18. Pecina, P., Dušek, O., Goeuriot, L., Hajič, J., Hlaváčová, J., Jones, G.J., Kelly, L., Leveling, J., Mareček, D., Novák, M., Popel, M., Rosa, R., Tamchyna, A., Urešová,

Z.: Adaptation of machine translation for multilingual information retrieval in the medical domain. Artificial Intelligence in Medicine (2014)
19. Koopman, B., Zuccon, G.: Relevation!: An open source system for information retrieval relevance assessment. In: Proceedings of the 37th annual international ACM SIGIR conference on research and development in information retrieval. (2014)
20. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
21. Robertson, S.E., Jones, S.: Simple, proven approaches to text retrieval. Technical Report 356, University of Cambridge (1994)
22. Thesprasith, O., Jaruskulchai, C.: Csku gprf-qe for medical topic web retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
23. Saleh, S., Pecina, P.: Cuni at the ShARe/CLEF eHealth Evaluation Lab 2014. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
24. Ozturkmenoglu, O., Alpkocak, A., Kilinc, D.: Demir at CLEF eHealth: The effects of selective query expansion to information retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
25. Dramé, K., Mougin, F., Diallo, G.: Query expansion using external resources for improving information retrieval in the biomedical domain. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
26. Shen, W., Nie, J.Y., Liu, X., Liui, X.: An investigation of the effectiveness of concept-based approach in medical information retrieval GRIUM @ CLEF2014eHealthTask 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
27. Thakkar, H., Iyer, G., Shah, K., Majumder, P.: Team IRLabDAIICT at ShARe/-CLEF eHealth 2014 Task 3: User-centered Information Retrieval system for Clinical Documents. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
28. Oh, H.S., Jung, Y.: A multiple-stage approach to re-ranking clinical documents. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
29. Ksentini, N., Tmar, M., Gargouri, F.: Miracl at CLEF 2014: eHealth information retrieval task. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
30. Verberne, S.: A language-modelling approach to user-centred health information retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
31. Claveau, V., Hamon, T., Grabar, N., , Maguer, S.L.: RePaLi participation to CLEF eHealth IR challenge 2014: leveraging term variation. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
32. Choi, S., Choi, J.: Exploring effective information retrieval technique for the medical web documents: SNUMedinfo at CLEFeHealth2014 Task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
33. Malagón, J.M.C., na López, M.J.M.: Laberinto at ShARe/CLEF eHealth Evaluation Lab. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
34. Yang, C., Bhattacharya, S., Srinivasan, P.: The University of Iowa at CLEF 2014: eHealth Task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
35. Wu, J., Huang, J.: York University at CLEF eHealth 2014: A Learning-to-Rank Approach for Medical Document Retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2014)
36. Goeuriot, L., Chapman, W., Jones, G.J.F., Kelly, L., Leveling, J., Salanterä, S.: Building realistic potential patients queries for medical information retrieval evaluation. In: Proceedings of the LREC workshop on Building and Evaluating Resources for Health and Biomedical Text Processing. (2014)

37. Goeuriot, L., Kelly, L., Leveling, J.: An analysis of query difficulty for information retrieval in the medical domain. In: Proceedings of SIGIR 2014 - Short papers track. (2014) to appear.