ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information Retrieval to Address Patients' Questions when Reading Clinical Reports

Lorraine Goeuriot[1], Gareth J.F. Jones[1], Liadh Kelly[1], Johannes Leveling[1], Allan Hanbury[2], Henning Müller[3], Sanna Salanterä[4], Hanna Suominen[5], and Guido Zuccon[6] *

[1] Dublin City University, Ireland, Firstname.Lastname@computing.dcu.ie
[2] Vienna University of Technology, Austria, hanbury@ifs.tuwien.ac.at
[3] HES–SO, Sierre, Switzerland, henning.mueller@hevs.ch
[4] University of Turku, Finland, sansala@utu.fi
[5] NICTA and The Australian National University, ACT, Australia, hanna.suominen@nicta.com.au
[6] The Australian e-Health Research Centre, CSIRO, QLD, Australia, guido.zuccon@csiro.au

Abstract. This paper presents the results of task 3 of the ShARe/CLEF eHealth Evaluation Lab 2013. 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 the discharge summaries and external resources such as medical ontologies on the IR effectiveness. The participants were allowed to submit up to seven runs, 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: before leaving the hospital, a patient receives a discharge summary. This describes the diagnosis and the treatment that they received in the hospital. The first task considered in CLEF eHealth aims at extracting names of disorders from the discharge summaries, while the second task requires normalisation and expansion of abbreviations and acronyms

* In alphabetical order, LG, GJFJ, LK & JL led Task 3; AH, HM, SS, HS & GZ were on the Task 3 organising committee.

present in the discharge summaries. The use case then postulates that, given the discharge summaries and the diagnosed disorders, patients often have questions regarding their health condition. The goal of the third task 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, queries, 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 [3], 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 characterised 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 Tracks aim at identifying patient cohorts from medical reports to recruit for user studies [4]. 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 hospitalisation 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 [5, 6]. 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.

The rest of this paper is organized as follows: Section 2 outlines the main evaluation campaigns on health IR. 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 analysed 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 [4]. 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.

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 topics, which were developed by medical experts, and the corresponding relevance information [13]. 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 training topics together with corresponding relevance assessment were released.

We describe in this section each part of the task dataset.

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

3.1 Document Collection

A large web crawl of health resources is used as the corpus for this task. The crawl contains about one million documents, which have been made available to CLEF eHealth through the Khresmoi project [14]. 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[8] as adhering to the HONcode principles[9] (approximately 60–70% of the collection), as well as other commonly used health and medicine websites such as Drugbank[10], Diagnosia[11] and Trip Answers[12]. 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.

3.2 Discharge Summaries

Novel methods to generate contextualised 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[13] (Multiparameter Intelligent Monitoring in Intensive Care, Version 2.5).

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
```

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

```
 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

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. Disorders have been identified within discharge summaries and linked to the matching UMLS (Unified Medical Language System) concept in the CLEF eHealth Task 1 [15].

Previous evaluation tasks in health IR have used MeSH entries (the MeSH ontology is contained in the UMLS meta-ontology) as queries (see Section 2). However, the queries considered by the task presented here are intended to be representative of real patients' information needs and statements. Thus the possibility of issuing concept-queries is discarded. Layperson queries tend to be short, with an average length less than two words. However, different patients can have different information needs associated with the same query statement. For example, a patient that receives a cancer diagnosis for the first time would have a different information need than a patient at a terminal cancer stage. This type of contextual information related to the patient history is contained in the discharge summary. Thus, the discharge summaries can be used for contextually focused generation of queries. The information in a discharge summary can then be used to determine the relevance of retrieved information to the specific user.

A query is generated for a given disorder and a discharge summary. To better structure the query generation process, patients' information needs have been grouped into three main scenarios:

1. the patient has a short-term disease, or has been hospitalised after an accident (little to no knowledge of the disorder, short-term treatment),
2. the patient has a chronic disease or a long-term disease that has *just* been diagnosed (little to no knowledge of the disorder, long-term treatment), and
3. the patient has a chronic or long-term disease, and this is the n-th diagnosis (potentially good knowledge of the disorder, long-term treatment).

Queries 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.

65 disorders have been randomly selected from the set of 1,006 disorders identified in the CLEF eHealth Task 1. For each disorder, a discharge summary containing the disorder itself has been randomly selected. Using the pairs of disorder and associated discharge summary, the experts have 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), and a narrative (expected content of the relevant documents).

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. 65 disorders have been selected (i.e. more than the targeted number of queries) because some disorders/queries may not be answerable using web pages from the document collection. During the query generation process, the experts manually removed disorders from the list of 65 that do not allow for realistic query generation. A real log containing queries issued by the general public on the HON website has also been used to exclude candidate queries which are unrealistic of the type of query that a patient would typically enter. For each query, an IR system that implements a standard BM25 weighting scheme [16] has been used to retrieve a shallow pool of documents. This has been used to assess whether a standard retrieval system could match at least one relevant document to a candidate query. Queries with no relevant documents retrieved in the shallow pool have been removed.

3.4 Relevance Assessment

Relevance assessment has been performed by domain experts and IR experts. We used the Relevation system[14] for collecting relevance assessments of documents contained in the assessment pools. Relevation is a system for performing relevance judgements for Information Retrieval system evaluation [17]. Documents

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

and queries can be uploaded to the system via the web interface; relevance judges can browse the uploaded documents and queries and provide their relevance assessments. The system is open source and based on Python's Django web framework. Relevation used a simple Model-View-Controller model that is designed for easy customisation and extension.

As we received many run submissions, we had to limit the pool depth and distributed the relevance assessment workload between medical experts and IR experts. The relevance assessment for these test queries has been conducted by six Finnish nursing professionals and five Australian nursing professionals or students in health sciences (domain experts), and three Irish, one Australian, and one Swedish senior researcher in clinical NLP (Natural Language Processing) and ML (Machine Learning), all technological experts. Each document was assessed by one person.

We pooled the top ten documents obtained from the participants' baseline runs, and both their top-priority run using discharge summaries and their top-priority run not using discharge summaries. This resulted in a pool of 6,391 documents. The relevance assessment was based on a four point scale, which is mapped to a binary scale. The graded relevance assessment yielded 0: 4,316, 1: 197, 2: 1,439, 3: 439 documents. The binary relevance assessments yielded 0: 4,513 non-relevant and 1: 1,878 relevant documents.

Thus, there have been 37.56 relevant documents per topic on average and 127.82 documents assessed per topic. Table 1 shows the relevance assessment coverage. (*) indicates runs for which results up to rank 10 were completely assessed.

Relevance assessments for the five training queries were formed based on pooled sets generated using the Vector Space Model [18] and Okapi BM25 [16]. Assessments for these five training queries were conducted by two Finnish nurses. Each document was assessed by one person.

Table 1. Coverage and minimum rank of first unassessed document for relevance assessments of runs. (*) indicates runs for which results up to rank 10 were completely assessed.

Run	# submitted runs	Assessed docs @ 10	Missing docs @ 10	Avg. minimum unassessed rank
1 (*)	9	100%	0	18.31
2 (*)	5	100%	0	14.75
3	5	92%	417	14.52
4	5	81%	940	14.02
5 (*)	9	100%	0	14.80
6	9	88%	1064	14.78
7	6	84%	955	14.76

4 Results

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

4.1 Participants

13 groups registered for task 3 and 9 groups submitted runs. The groups are from 5 countries and 4 continents: Asia (2), Australia (2), USA (4), Europe (1). The groups are listed in Table 2. Although CLEF has mainly been the focus of European groups, only one group from Europe submitted runs to this lab.

Table 2. Participants for task 3.

Name	Country	# submitted runs	Run prefix
Team-OHSU	(USA)	3	ohsu
Team-UCSC.KC&RA	(USA)	4	TeamKC
Team-Mayo	(USA)	7	TeamMayo
Team-UTHealth_CCB-3	(USA)	4	UTHealth.CCB
Team-UOG	(UK)	4	uogTr
Team-AEHRC	(Australia)	4	teamAEHRC
Team-QUT	(Australia)	6	QUT-TOPSIG
Team-SNUBMedinfo	(Republic of Korea)	7	MEDINFO
Team-THCIB	(China)	7	THCIB

Teams submitted in total 48 runs, including: 9 baseline runs, 15 runs using discharge summaries (from 5 teams), and 24 runs not using them.

4.2 Evaluation Metrics

We examined all documents in runs 1, 2 and 5 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. Table 1 provides details on the coverage of assessments for each run up to rank 10.

Performance metrics are computed with the standard *trec_eval* tool[15] using the following options:

- trec_eval -c -M1000
- trec_eval -c -M1000 -m ndcg_cut

We are aware that the performance metrics for other runs might be unreliable compared to that of runs 1, 2, and 5. 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 experiment, which is based on the BM25 retrieval model. This experiment does not incorporate any domain-specific adaptations. We used the JSoup library[16] to extract the textual content from the web pages and applied standard normalization (e.g. replacing named HTML characters) and spelling correction based on a fixed list of frequent spelling errors. This approach has been employed before for experiments on medical records retrieval [19] and was shown to represent a very strong baseline.

This baseline system uses the BM25 retrieval model and standard blind relevance feedback for BM25 [16]. The system uses a standard stop-word list containing the Okapi stop-words (222 stop-words) for stop-word removal.

The baseline system performs only two types of document preprocessing: character normalization (i.e. mapping characters with diacritical marks to the equivalent characters without) and word normalization (e.g. correcting frequent spelling errors). Spelling correction is based on a list of 9533 spelling errors from medical documents [19], which was added to a list of 4192 frequent spelling errors compiled from Wikipedia. During indexing, misspelled words are replaced with their corrections from this list.

We generated two baseline experiments: one with standard retrieval using the BM25 model (BM25_baseline) and a second experiment using blind relevance feedback (BM25_FB_baseline). For blind relevance feedback, the top T terms from the top ranked D documents are added to the query to retrieve the final result set of documents. For our baseline experiments, we used $T = 10$ and $D = 10$.

4.4 Evaluation Results

The official results for all submitted runs and for our baseline experiments are shown in Table 3.

Comparing the participants' results wrt. P@10 and NDCG@10, we found that for 4 teams (ohsu, THCIB, teamAEHRC, uog) run 5 (using discharge summaries) achieves the best P@10, while for 2 teams (QUT-TOPSIG, TeamMayo),

[15] http://trec.nist.gov/trec_eval/
[16] http://jsoup.org/

Table 3. Evaluation in Task 3. Results where the baseline is significantly better (or results significantly worse than the baseline) are indicated by "*" (Wilcoxon test with 95% confidence). No submitted results are significantly better than the baseline. BM25 is the baseline provided by the organisers, using the BM25 retrieval model and relevance feedback (BM25_FB). The best P@10 values for each team are *emphasised*.

Run ID	P@5	P@10	NDCG@5	NDCG@10	MAP	rel_ret
TeamMayo.1.3	0.4800	0.4720	0.4370	0.4408	0.3040	1619
TeamMayo.2.3	0.4960	*0.5180*	0.4391	0.4665	0.3108	1673
TeamMayo.3.3	0.5280	0.4880	0.4742	0.4584	0.2900	1689
TeamMayo.4.3	0.5240	0.4820	0.4837	0.4637	0.2967	1689
TeamMayo.5.3	0.5120	*0.5040*	0.4645	0.4618	0.3061	1689
TeamMayo.6.3	0.5160	0.4940	0.4639	0.4579	0.2953	1689
TeamMayo.7.3	0.4920	0.4700	0.4348	0.4332	0.2981	1689
teamAEHRC.1.3	0.4440	*0.4540*	0.3814	0.3980	0.2462	1286
teamAEHRC.5.3	0.4560	*0.4840*	0.3957	0.4226	0.2732	1495
teamAEHRC.6.3	0.4440	0.4240	0.4117	0.3993	0.2442	1477
teamAEHRC.7.3	0.2080	0.2200*	0.1926	0.1984	0.1589	1425
MEDINFO.1.3	0.4600	*0.4800*	0.4189	0.4377	0.3131	1663
MEDINFO.2.3	0.4040	0.3980*	0.3467	0.3546	0.2454	1609
MEDINFO.3.3	0.4280	0.4040*	0.3703	0.3639	0.2584	1622
MEDINFO.4.3	0.4200	*0.4060**	0.3667	0.3691	0.2601	1618
MEDINFO.5.3	0.3960	0.4040*	0.3407	0.3561	0.2426	1609
MEDINFO.6.3	0.3880	0.3600*	0.3326	0.3284	0.2343	1605
MEDINFO.7.3	0.3560	0.3480*	0.3061	0.3075	0.2174	1551
uogTr.1.3	0.4240	*0.4360*	0.3708	0.3807	0.2438	1005
uogTr.5.3	0.4280	*0.4400*	0.3663	0.3840	0.2429	983
uogTr.6.3	0.4120	0.4040	0.3470	0.3528	0.2186	978
uogTr.7.3	0.3640	0.3500*	0.3229	0.3207	0.1923	961
THCIB.1.3	0.4360	0.3960*	0.3923	0.3716	0.1028	198
THCIB.2.3	0.4440	0.3980	0.4026	0.3808	0.1106	199
THCIB.3.3	0.4400	0.4020	0.3966	0.3811	0.1031	201
THCIB.4.3	0.3160	0.3080*	0.2800	0.2910	0.0786	154
THCIB.5.3	0.4800	*0.4200*	0.4352	0.4044	0.1217	210
THCIB.6.3	0.4560	*0.4140*	0.4100	0.3904	0.1155	207
THCIB.7.3	0.3360	0.3080*	0.2984	0.2928	0.0729	154
teamKC.1.3	0.4040	*0.4040**	0.3587	0.3637	0.2666	1646
teamKC.2.3	0.0720	0.0600*	0.0589	0.0548	0.0178	217
teamKC.3.3	0.2040	0.1920*	0.1759	0.1765	0.1590	1465
teamKC.4.3	0.2520	0.2320*	0.2133	0.2062	0.1634	1433
teamKC.5.3	0.0680	0.0580*	0.0586	0.0549	0.0197	250
teamKC.6.3	0.3440	*0.3640**	0.3144	0.3281	0.2270	1561
UTHealth.CCB.1.3	0.3920	*0.3740*	0.3444	0.3406	0.1482	458
UTHealth.CCB.5.3	0.2600	0.2540*	0.2681	0.2587	0.0953	296
UTHealth.CCB.6.3	0.2760	*0.2560**	0.2384	0.2337	0.1124	337
UTHealth.CCB.7.3	0.1680	0.1460*	0.1442	0.1368	0.0546	204
QUT-TOPSIG.1.3	0.3680	*0.3620**	0.3376	0.3419	0.2014	1492
QUT-TOPSIG.2.3	0.3680	*0.3640**	0.3281	0.3368	0.2009	1492
QUT-TOPSIG.3.3	0.3200	0.3320*	0.2808	0.2948	0.1872	1458
QUT-TOPSIG.4.3	0.0720	0.0560*	0.0669	0.0617	0.0342	450
QUT-TOPSIG.5.3	0.3200	0.3320*	0.2808	0.2944	0.1859	1458
QUT-TOPSIG.6.3	0.0960	0.0900*	0.0876	0.0819	0.0745	1195
ohsu.1.3	0.2800	*0.2300**	0.2719	0.2436	0.0953	625
ohsu.5.3	0.2840	*0.2600**	0.2350	0.2344	0.0999	333
ohsu.6.3	0.1920	0.1620*	0.1895	0.1706	0.0816	461
BM25_FB	0.4840	0.4860	0.4205	0.4328	0.2945	1636
BM25	0.4520	0.4700	0.3979	0.4169	0.3043	1651

run 2 (not using summaries) achieves the best P@10. For 3 teams (TeamKC, MEDINFO, UTHealth), the baseline runs (run 1) performs best.

Only 4 runs (from the same group) outperform our baseline experiment using standard blind relevance feedback wrt. P@10. Runs from two groups outperform the baseline experiment wrt. NDCG@10. This illustrates that our baseline system based on BM25 is a very strong baseline.

5 Approaches Used

We describe in this section the approaches used by each team, and summarize findings from their analysis.

Team THCIB [20] implemented ad hoc retrieval using Lucene for indexing and retrieval and HITS for ranking. They submitted 7 runs. They investigated query annotations with UMLS, and the use of various fields of the queries. They also used query expansion based on the concepts and/or the discharge summaries, as well as concept-based re-ranking. Their best run scores 0.42 P@10 (run 5), and uses their baseline system, with query expansion without the discharge summaries on the *title* and *desc* topic fields.

Team TOPSIG [21] used the TopSig open source tool, which implements signature-based approaches for information retrieval. Six runs were submitted, using various parts of the topic files as queries, and using the text of the discharge reports for query refinement. Submissions used TopSig normally for query-only runs and with a new 'refine' mode for discharge summary runs. This means the query is used for searching the collection and the results are then re-ranked using the discharge summaries, similarly to how blind feedback works. This produced much better results than simply using the discharge summaries in the original query (P@10 of 0.36), which tend to be too noisy to produce effective results.

Team AEHRC [22] used a Dirichlet-smoothed language modelling as provided by the Terrier system for retrieval. They experimented with setting prior probabilities depending on document readability and authorativeness information derived from a static list of web sites. They also explored spelling correction (using Google) and acronym expansion. They submitted 4 runs, and did not use the discharge summaries. Their best result (P@10 of 0.48) is obtained with the baseline and the acronym expansion and spelling correction in the query.

Team KC [23] approach is based on statistical topic models. Their baseline used Language models for retrieval, with Indri index engine. They submitted 6 runs, exploring query expansion through the extraction and selection of noun phrases from documents and discharge summaries, using probabilistic topic modelling and language models. They scored their best result (P@10 of 0.40) with the baseline.

Team Mayo [24] used a Query-Likelihood Model as their baseline, and further runs using a Markov Random Field Model. Query expansion was carried out using a Mixture of Relevance Models and using the MeSH ontology. They also investigated concept-based re-ranking using medical concepts identified in the discharge summaries, and weighting concepts using attributes such as negation or semantic groups. Their run # 2 got the best results (P@10 of 0.52), using a Markov random field to model the dependency between query terms, expansion of query terms through 4 medical/genomics collections and a concept-space search ranking (based on query and discharge summary).

Team SNUMEDINFO [25] used unigram language model with Dirichlet prior smoothing on indri search engine as a baseline. Their runs without discharge summaries are using passage based language model, combining max-scoring passage-based relevance score with unigram language model score, with different weighting parameters. The following runs used lexical query expansion with UMLS concepts and preferred terms. They also explored the use of different fields of the query. Their best results are obtained with the baseline (P@10 of 0.48).

Team OHSU [26] submission included runs from two different retrieval systems. One set used the Lucene Vector Space Retrieval model and extensions which made use of MetaMap for query expansion. The second used novel statistical language modelling techniques. Valuable insights for future system improvement were gained, such as insights into selective indexing due to the large amount of noise in web page content, and the need for sophisticated query parsing and expansion techniques. They obtained best results with run #5, using a language model to attempt to find the documents whose word distributions would have been most likely to generate the query, together with an adaptive perplexity threshold.

Team UThealth [27] submission contained runs using the vector space model and the semantic vector space model. Best performance was obtained using the vector space model. They also explored the use of different fields of the topics as queries. Their best result was obtained with the baseline (P@10 of 0.37).

Team UOG [28] used the Terrier information retrieval framework. Divergence from Randomness and pseudo relevance feedback were employed on the retrieval of medical web pages. They also investigated query expansion using corpus of MEDLINE abstracts or Wikipedia collection. Their best run (P@10 of 0.44) used the baseline system and pseudo relevance feedback.

The best result overall was obtained by team Mayo, using a retrieval model and external resources for query expansion, as well as re-ranking based on concepts from the query and the discharge summary. However, this team, together with Topsig, are the only teams having improved their baseline with the use of the discharge summaries, and in both cases for re-ranking purposes.

Concept re-ranking has also been used by team THCIB and given their best results. For five teams, the best scores are obtained with their baseline (coupled with pre-processing such as acronym expansion and spelling correction for team AEHRC). Combination of methods and external resources such as those employed by team Mayo seem to be efficient, but the overall results here show that research still needs to be conducted to make the best out of the external resources.

6 Conclusions

In this first year of the ShARe/CLEF eHealth2013 evaluation lab Task 3, there was strong take-up in the community with 9 groups submitting runs to the task. The challenge of developing retrieval techniques for layperson medical queries proved difficult. Overall, BM25 is a very strong baseline, which proved hard for teams to beat. BM25 with relevance feedback, as opposed to standard BM25, proved to be the strongest baseline. Here P@10 of 0.4860 was obtained, and NDCG@10 of 0.4328. One team, Mayo, had four runs which performed better than BM25 for P@10. P@10 achieved for these runs ranged from 0.5180 – 0.4880. Of these runs, the best performing used information in the discharge summaries. The only other team that improved over the BM25 baseline was SNUBME for their baseline run, where NDCG@10 of 0.4377 was achieved.

Despite the results obtained being mostly lower than the baseline in this first year of the task, many valuable insights into retrieval technique development for the domain were gained. This forms a good basis for future exploration in the domain in general, and specifically for further technique development for the 2014 CLEF eHealth Task 3 lab.

Given the success of this year 1 of the task, we anticipate even more interest in next year's campaign. In the second year of the task, we will seek to remove noise from the document collection. This year, as task participants were informed, there were some non-English web pages in the collection and duplicate web pages. That is there were occurrences in the document collection of the same web page, with the same URL but different unique identifier. On top of this, we also identified several occurrences of the same web pages with different URL prefixes. This occurs for example, when a dropdown menu is expanded. We will also explore the possibility of removing or highlighting such duplication in the collection. In next year's campaign we also intend exploring new query generation techniques and means to improve the relevance assessment workflow. In query generation for example, we will look at using the main disease in the discharge summary the query (information need) stems from instead of using a randomly selected disease from within the discharge summary. The goal here being to increase the relevance of using discharge summaries.

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, and NICTA, funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. The relevance judgements were in part funded by the ESF project ELIAS. We acknowledge the time given to perform the relevance assessment task. We want to thank the following individuals: Maricel Angel (NICTA, Australia), Rmi Bois (Dublin City University), Riitta Danielsson-Ojala (University of Turku, Finland), Lotta Kauhanen (University of Turky, Finland), Hugo Mougard (Dublin City University), Laura-Maria Murtola (University of Turku, Finland), Heidi Parisod (University of Turku, Finland), Josh Robertson (Dublin City University, Ireland), Eriikka Siirala (University of Turku, Finland), Timothy Sladden (University of Queensland, Australia), Thomas Souchen (AEHRC, Australia), Sumithra Velupillai (Stockholm University, Sweden).

References

1. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Mowery, D., Leveling, J., Goeuriot, L., Kelly, L., Martinez, D., Zuccon, G.: ShARe/CLEF eHealth Evaluation Lab 2013: Three shared tasks on natural language processing and machine learning to make clinical reports easier to understand for patients. In: CLEF 2013. Lecture Notes in Computer Science (LNCS), Springer (2013)
2. Fox, S.: Health topics: 80% of internet users look for health information online. Technical report, Pew Research Center (February 2011)
3. Hanbury, A.: Medical information retrieval: an instance of domain-specific search. In: Proceedings of SIGIR 2012. (2012) 1191–1192
4. Voorhees, E.M., Tong, R.M.: Overview of the TREC 2011 medical records track. In: Proceedings of TREC, NIST (2011)
5. 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)
6. 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)
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. 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)
14. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: MIE village of the future. (2012)
15. S. Pradhan, N. Elhadad, B.S.D.M.L.C.H.W.C.G.S.: Task 1: Share/clef ehealth. In: Proceedings of the CLEF conference. (2013)
16. Robertson, S.E., Jones, S.: Simple, proven approaches to text retrieval. Technical Report 356, University of Cambridge (1994)
17. Koopman, B., Zuccon, G.: Relevation! an open source system for information retrieval relevance assessment. arXiv preprint (2013)
18. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
19. Leveling, J., Goeuriot, L., Kelly, L., Jones, G.J.F.: DCU@TRECMed 2012: Using ad-hoc baselines for domain-specific retrieval. In: Proceedings of TREC 2012, NIST (2012)
20. Zhong, X., Xia, Y., Xie, Z., Na, S., Hu, Q., Huang, Y.: Concept-based medical document retrieval: THCIB at CLEF eHealth lab 2013 task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2013)
21. Chappell, T., Geva, S.: Working notes for topsig at share/clef ehealth 2013. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2013)
22. Zuccon, G., Koopman, B., Nguyen, A.: Retrieval of health advice on the web: AEHRC at ShARe/CLEF eHealth evaluation lab task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2013)
23. Barajas, K.C., Akella, R.: Incorporating statistical topic models in the retrieval of health care documents. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2013)
24. Zhu, D., Wu, S., James, M., Carterette, B., Liu, H.: Using discharge summaries to improve information retrieval in clinical domain. In: Proceedings of the ShARe/-CLEF eHealth Evaluation Lab. (2013)
25. Choi, S., Choi, J.: SNUMedinfo at CLEFeHealth2013 task 3. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2013)
26. Bedrick, S., Sheikhshabbafghi, G.: Lucene, metamap, and language modeling: OHSU at CLEF eHealth 2013. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2013)
27. Zhang, Y., Cohen, T., Jiang, M., Tang, B., Xu, H.: Evaluation of vector space models for medical disorders information retrieval. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2013)

28. Limsopatham, N., Macdonald, C., Ounis, I.: University of glasgow at CLEF 2013: Experiments in eHealth task 3 with terrier. In: Proceedings of the ShARe/CLEF eHealth Evaluation Lab. (2013)