Overview of the ShARe/CLEF eHealth Evaluation Lab 2013

Hanna Suominen[1], Sanna Salanterä[2], Sumithra Velupillai[3], Wendy W. Chapman[4], Guergana Savova[5], Noemie Elhadad[6], Sameer Pradhan[5], Brett R. South[7], Danielle L. Mowery[8], Gareth J. F. Jones[9], Johannes Leveling[9], Liadh Kelly[9], Lorraine Goeuriot[9], David Martinez[10], and Guido Zuccon[11⋆]

[1] NICTA and The Australian National University, ACT, Australia, Hanna.Suominene@nicta.com.au. Corresponding author.
[2] University of Turku, Finland, sansala@utu.fi
[3] DSV Stockholm University, Sweden, sumithra@dsv.su.se
[4] University of California, San Diego, CA, USA, wwchapman@ucsd.edu
[5] Harvard University, MA, USA, Firstname.Lastname@childrens.harvard.edu
[6] Columbia University, NY, USA, noemie@dbmi.columbia.edu
[7] University of Utah, UT, USA, brett.south@hsc.utah.edu
[8] University of Pittsburgh, PA, USA, dlm31@pitt.edu
[9] Dublin City University, Ireland, Firstname.Lastname@computing.dcu.ie
[10] NICTA and The University of Melbourne, VIC, Australia, David.Martinez@nicta.com.au
[11] The Australian e-Health Research Centre, CSIRO, QLD, Australia, Guido.Zuccon@csiro.au

Abstract. Discharge summaries and other free-text reports in healthcare transfer information between working shifts and geographic locations. Patients are likely to have difficulties in understanding their content, because of their medical jargon, non-standard abbreviations, and ward-specific idioms. This paper reports on an evaluation lab with an aim to support the continuum of care by developing methods and resources that make clinical reports in English easier to understand for patients, and which helps them in finding information related to their condition. This ShARe/CLEFeHealth2013 lab offered student mentoring and shared tasks: identification and normalisation of disorders (1a and 1b) and normalisation of abbreviations and acronyms (2) in clinical reports with respect to terminology standards in healthcare as well as information retrieval (3) to address questions patients may have when reading clinical reports. The focus on patients' information needs as opposed to the specialised information needs of physicians and other healthcare workers was the main feature of the lab distinguishing it from previous shared tasks. De-identified clinical reports for the three tasks were from US intensive care and originated from the MIMIC II database. Other text documents for Task 3 were from the Internet and originated from

⋆ In alphabetical order, HS, SS & SV co-chaired the lab; WWC led Tasks 1 & 2, GS, NE & SP as the leaders of Task 1 and BRS & DLM as the leader of Task 2; GJFJ, JL, LK & LG led Task 3; and DM & GZ were the leaders of result evaluations.

the Khresmoi project. Task 1 annotations originated from the ShARe annotations. For Tasks 2 and 3, new annotations, queries, and relevance assessments were created. 64, 56, and 55 people registered their interest in Tasks 1, 2, and 3, respectively. 34 unique teams (3 members per team on average) participated with 22, 17, 5, and 9 teams in Tasks 1a, 1b, 2 and 3, respectively. The teams were from Australia, China, France, India, Ireland, Republic of Korea, Spain, UK, and USA. Some teams developed and used additional annotations, but this strategy contributed to the system performance only in Task 2. The best systems had the F1 score of 0.75 in Task 1a; Accuracies of 0.59 and 0.72 in Tasks 1b and 2; and Precision at 10 of 0.52 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, Evaluation, Medical Informatics, Test-set Generation, Text Classification, Text Segmentation

1 Introduction

Discharge summaries transfer information between working shifts and geographical locations. They are written or dictated by a physician, nurse, therapist, specialist, or other clinician responsible for patient care to describe the course of treatment, the status at release, and care plans. Their primary purpose is to support the care continuum as a handover note between clinicians, but they also serve legal, financial, and administrative purposes. In several countries these documents are regulated by law. For example, in Sweden, the Patient Data Law 255/2008 and in Finland, the Statute 298/2009 on Patient Documents state that in order to ensure good care, clinical documents must cover all necessary information and adequately detail the patient's conditions, care, and recovery. This legislation also stipulates that the documents must be explicit, comprehensive, and include only generally well-known, accepted concepts and abbreviations.

However, the law and practice differ substantially [1, 2]. The patient and her next of kin are likely to have difficulties in understanding this simple example sentence from a US discharge: "*AP: 72 yo f w/ ESRD on HD, CAD, HTN, asthma p/w significant hyperkalemia & associated arrythmias.*" After expanding the abbreviations and acronyms as well as correcting the misspellings, they are much more likely to understand that this sentence belongs to the description of the patient's *active problem*. It tells that the patient is a *72 year old female with dependence on hemodialysis, coronary heart disease, hypertensive disease, and asthma*. Her current medical problem (i.e., *presenting problem*) is *significant hyperkalemia and associated arrhythmias*. An improved understanding of related concepts in discharge summaries can be achieved by normalising all health conditions to standardised, computer-processable language. In SNOMED-CT, the

CUIs *C0003811*, *C0004096*, and *C0020461* correspond to synonyms of arrhythmia, asthma, and hyperkalemia, respectively.[12]

The patient's and her next-of-kin's understanding of health conditions can be supported not only by these expansions, corrections, and normalisations, but also by linking the words to a patient-centric search on the Internet. Already without electronic linkage with discharge summaries, nearly 70 per cent of search engine users in the USA in 2012 searched for information about health conditions [3]. In 2007, nearly 47 per cent of Europeans considered the Internet as an important source of health information [4] and over 42 per cent of Australian searches were related to health and medical information [5]. The search engine could, for example, link hyperkalemia and its synonyms to definitions in Wikipedia, Consumer Health Vocabulary, and other patient-friendly sources.[13] 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 also assess the reliability of information (e.g., guidelines by healthcare service providers vs. uncurated but insightful experiences on discussion forums).

This paper presents an overview of the ShARe/CLEFeHealth2013 evaluation lab[14] to address these approaches in making clinical text easier to understand and targeting patients' information needs in search on the Internet. The novel lab aimed to develop processing techniques and data for these approaches and an evaluation setting that includes statistical metrics of correctness and end-user engagement by asking nurses and laypeople to represent patients' preferences in expansions, normalisations, and search. It offered a mentoring track for graduate students working on related fields and shared tasks on NLP and ML: identification and normalisation of disorders (1a and 1b) [6] and normalisation of abbreviations and acronyms (2) [7] in clinical reports with respect to terminology standards in healthcare as well as IR (3) [8] to address questions patients may have when reading clinical reports[15]. This attracted 34 teams to submit 113 systems[16]; 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 in CLEF on 23–26 Sep 2013.

2 Background

For over forty years, NLP and other techniques based on computational linguistics and ML have been recognised as ways to automate text analysis in health-

[12] Systematized Nomenclature of Medicine Clinical Terms, Concept Unique Identifiers

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

[14] `http://nicta.com.au/business/health/events/clefehealth_2013`, Shared Annotated Resources, `http://clinicalnlpannotation.org`, and Conference and Labs of the Evaluation Forum

[15] Natural Language Processing, Machine Learning, and Information Retrieval

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

care. PubMed[17] returns 12,860 references, including pioneering studies [9–12] and recent reviews [13–18]. Some techniques have progressed from research to use in practice. As US examples, MedLEE[18] used in the New York Presbyterian Hospital normalises patient records to UMLS[19] [19] and Autocoder at the Mayo Clinic in Rochester assigns diagnosis codes to patient records, reducing workload by 80 per cent [20]. However, the development and progress has been substantially hindered, but shared tasks address these barriers [21]. The barriers can be classified as lack of access to shared data for system researh, development and evaluation; insufficient common conventions and standards for data, technologies, and evaluations; the formidability of reproducibility; limited collaboration; and lack of user-centered development and scalability.

The first shared tasks related to clinical NLP were in TREC[20]. The 2000 Filtering Track [22] focused on building user profiles to separate relevant and irrelevant documents. Data contained around 350,000 abstracts from the MEDLINE database over five years, manually created topics, and a topic set based on the standardised MeSH.[21] The Genomics Track [23] had in 2003–2007 annual IR tasks on genomics data in biomedical papers and clinical reports. The tasks ranged from ad-hoc IR to classification, passage IR, and entity-based question-answering. The Medical Records Track [24] in 2011 and 2012 aimed to develop an IR technique for finding patient cohorts that are relevant to a given criteria for recruitment as populations in comparative effectiveness studies. Their data consisted of de-identified medical records, queries that resemble eligibility criteria of clinical studies, and associated relevance assessments.

In 2005, ImageCLEFmed[22] [25, 26] introduced annual tasks on accessing to biomedical images in papers and on the Internet. In 2005–2013, it targeted language-independent techniques for annotating images with concepts; multimodal IR combining visual and textual features; and multilingual IR techniques.

In 2006, i2B2[23] [27] began its tasks on clinical NLP: text de-identification and identification of smoking status in 2006; recognition of obesity and co-morbidities in 2008; medication information extraction in 2009; concept, assertion, and relation recognition in 2010; co-reference analysis in 2011; and temporal-relation analysis in 2012. Data originated from the USA, were in English, and included approximately 1,500 de-identified discharge summaries with their annotations.

Medical NLP Challenges[24] by the Computational Medicine Center in 2007 [28] and 2011 [29] addressed automated diagnosis coding of radiology reports and classifying the emotions found in suicide notes. In 2007, 1,954 de-identified

[17] the query of *(natural language processing) OR (text mining)* on 27 Jun 2013
[18] Medical Language Extraction and Encoding System
[19] Unified Medical Language System
[20] Text Retrieval Conference, `http://trec.nist.gov/data/filtering.html`, `http://ir.ohsu.edu/genomics/`, and `http://trec.nist.gov/data/medical.html`
[21] Medical Literature Analysis and Retrieval System Online and Medical Subject Headings
[22] `http://ir.ohsu.edu/image/`
[23] Informatics for Integrating Biology and the Bedside, `https://www.i2b2.org/`
[24] `http://computationalmedicine.org/challenge/`

radiology reports in English from a US radiology department for children were used. In 2011, over 1,000 suicide notes in English were used.

In 2013, the Health Design Challenge[25] challenged to re-imagine the visuals and layout of health/medical records. The purpose was to make the records more usable by and meaningful to patients, their families, and others who take care of them. The challenge was motivated by the continuum of care but did not address NLP and ML. Over 230 teams submitted their designs. The winning designs were announced in Jan 2013 and are showcased on the Internet.

In Nov 2012 – Feb 2013, NTCIR ran MedNLP[26] on information extraction from simulated medical reports in Japanese. It had text de-identification, complaint/diagnosis recognition, and open tasks.

Targeting patients' information needs through NLP, ML and IR is important, novel, and difficult. Meeting these needs is critical because of the empowering effects the right information and the negative effects missing or incorrect information may have on health outcomes. The focus on patients' and next-of-kins' information needs as opposed to the specialised information needs of healthcare workers is the main distinguishing feature of the ShARe/CLEFeHealth 2013 evaluation lab compared to previous shared tasks. This is, however, technically more difficult, as they represent a wider and more heterogeneous subject population. The variance in, for example, their health profiles, health knowledge, abilities to interpret health information, computer skills, and search queries is greater [30].

3 Materials and Methods

3.1 Text Documents

For Tasks 1–3, de-identified clinical reports were from US intensive care and originated from the ShARe corpus[27] which has added layers of annotation over the clinical notes in the version 2.5 of the MIMIC II database[28]. The corpus consisted of discharge summaries and electrocardiogram, echocardiogram, and 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[29], 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.

[25] `http://healthdesignchallenge.com`

[26] NII Test Collection for IR Systems, `http://mednlp.jp/medistj-en`

[27] `https://www.clinicalnlpannotation.org`

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

[29] 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`.

For Task 3, a large crawl of health resources on the Internet was used. It contained about one million documents [31] and originated from the Khresmoi project[30]. The crawled domains were predominantly of 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.[31] Documents consisted of pages on a broad range of health topics and 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.[32]

3.2 Human Annotations, Queries, and Relevance Assessments

For Task 1, annotation of disorder mentions in clinical reports was carried out as part of the ongoing ShARe project. For this task in the evaluation lab, the focus was on the annotation of disorder mentions only. As such, there were two parts to the annotation: identifying a span of text as a disorder mention and mapping the span to a UMLS CUI. Each note was annotated by two professional coders trained for this task, followed by an open adjudication step. UMLS[33] represented over 130 lexicons/thesauri with terms from a variety of languages. It integrated resources used world-wide in clinical care, public health, and epidemiology. It also provided a semantic network in which every concept is represented by its CUI and is semantically typed [32]. A disorder mention was defined as any span of text which can be mapped to a concept in SNOMED-CT and which belongs to the Disorder semantic group.[34] A concept was in the Disorder semantic group if it belonged to one of the following UMLS semantic types: Congenital Abnormality; Acquired Abnormality; Injury or Poisoning; Pathologic Function; Disease or Syndrome; Mental or Behavioral Dysfunction; Cell or Molecular Dysfunction; Experimental Model of Disease; Anatomical Abnormality; Neoplastic Process; and Signs and Symptoms. The annotations covered about 181,000 words.

For Task 2, a gold standard of acronyms and abbreviations normalised to CUIs from the UMLS was developed. It was generated in the following three phases: First, one Australian and nine Finnish nursing professionals as well as an Australian senior researcher in clinical NLP and ML were trained for the task using annotation guidelines and the eHOST[35]annotation tool [33]; provided reports

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

[31] 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`

[32] HyperText Markup Language and Uniform Resource Locators

[33] `https://uts.nlm.nih.gov/home.html`

[34] Note that this definition of Disorder semantic group did not include the Findings semantic type, and as such differed from the one of UMLS Semantic Groups, available at `http://semanticnetwork.nlm.nih.gov/SemGroups`.

[35] extensible Human Oracle Suite of Tools, `http://orbit.nlm.nih.gov/resource/ehost-extensible-human-oracle-suite-tools`

from Task 1 with disorder annotations; and instructed to span clinical acronym and abbreviations in the clinical reports. When possible, a spanned concept was assigned one CUI from the UMLS; otherwise, it was assigned "CUI-less". Second, Phase 1 annotations were adjudicated by a US biomedical informatician as the silver standard. Third, Phase 2 annotations were adjudicated by a US biomedical informatician certified as a respiratory therapist creating the final gold standard. The Phase 3 annotations covered approximately 7,500 abbreviations in total.

For Task 3, queries and the respective result sets and relevance assessments were associated with the text documents [34]. Two Finnish nursing professionals created 55 queries from highlighted disorders identified in Task 1 (a manually extracted set). They also generated a mapping between queries and the matching clinical report in Task 1. This was provided to the participants but they were also free to use the clinical report, if they had access to them. Relevance assessments were performed by domain experts and technological experts using the Relevation system[36] [35] for collecting relevance assessments of documents contained in the assessment pools. Documents and queries were uploaded to the system via the Internet interface; judges could browse the uploaded documents and queries and provide their relevance assessments. The domain experts included six Finnish nursing professionals and five Australian nursing professionals or students in health sciences. The technological experts included three Irish, one Australian, and one Swedish senior researcher in clinical NLP and ML. Assessments compared the query and its mapping to the content of the retrieved document on a four-point scale (Fig. 1). The relevance of each document was assessed by one expert. The 55 queries were divided between training and testing. Assessments for the 5 training queries were performed by the same two Finnish nursing professionals who generated the queries. As we received 48 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 33 test queries were assessed by the domain experts and the remaining 17 test queries by the technological experts.

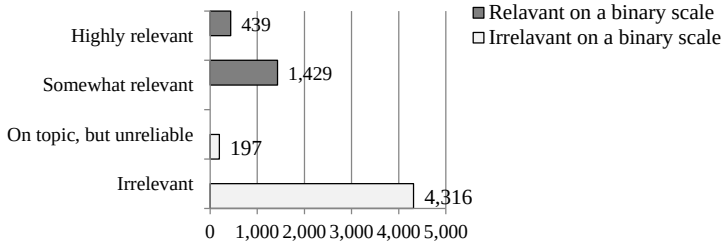

Fig. 1. Distribution of the relevance assessments on 4-point and binary scales

[36] `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

3.3 Evaluation Methods

The following evaluation criteria were used: correctness in identification of the character spans of disorders (1a), correctness in mapping disorders to SNOMED-CT codes (1b), correctness in mapping pre-identified acronyms/abbreviations to UMLS codes (2), and relevance of the retrieved documents to patients or their representatives.

In Tasks 1a, 1b, and 2, each participating team was permitted to upload the outputs of up to two systems. Task 1b was optional for Task 1 participants. Teams were allowed to use additional annotations in their systems, but this counted towards the permitted systems; systems that used annotations outside of those provided for Tasks 1 and 2 were evaluated separately. In Task 3, 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).

Teams received training and test datasets in Feb–May, 2013. The evaluation for all tasks was conducted using the blind, withheld test data (reports for Tasks 1 and 2 and queries for Task 3). Teams were asked to stop development as soon as they downloaded the test data. The training set (test set) was released on 15 Feb (17 Apr), 21 Mar (1 May), and 25 Mar – 15 Apr (24 Apr) for Tasks 1, 2, and 3, respectively. Outputs for the test set were due by (evaluation results were announced to the participants on) 24 Apr (14 May), 8 May (17 May), and 1 May (1 Jun) to Tasks 1, 2, and 3, respectively.

In Tasks 1 and 2, participants were provided a training set containing clinical text as well as pre-annotated spans and named entities for disorders (Tasks 1a and 1b) or acronyms/abbreviations (Task 2). For Task 1a, participants were instructed to develop a system that predicts the spans for disorder named entities. For Tasks 1b and 2, participants were instructed to develop a system that predicts the SNOMED-CT (Task 1b) or UMLS (Task 2) CUI code (or CUI-less) for unknown pre-annotated spans. The outputs needed to follow the annotation format. The corpus of reports was split into 200 training and 100 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 highest priority output from the runs 2–4, and the highest priority output from the runs 5–7 were pooled with duplicates removed. This resulted in a pool of 6,391 documents (Fig. 1). Pooled sets for the training queries were created by merging the top 30 ranked documents returned by the two IR models (Vector Space Model [36] and BM25 [37]) and removing duplicates.

The system performance was evaluated agaist the criteria by using the F1 score in Task 1a, Accuracy in Tasks 1b and 2, and Precision at 10 in Task 3. We relied on non-parametric statistical significance tests called random shuffling [38] in Tasks 1 and 2, and the Wilcoxon test [39] in Task 3 to better compare the measure values for the systems and benchmarks.

In Task 1a, 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 Tasks 1b and 2, the Accuracy was defined as the number of pre-annotated spans with correctly generated code divided by the total number of pre-annotated spans. In both tasks, the Exact Accuracy and Relaxed Accuracy were measured. In the Exact Accuracy for Task 1b, *total* was defined as the total number of gold standard named entities. In this case, the system was penalised for incorrect code assignment for annotations that were not detected by the system. In the Relaxed Accuracy for Task 1b, *total* was defined as the total number of named entities with strictly correct span generated by the system. In this case, the system was only evaluated on annotations that were detected by the system. In the Exact Accuracy for Task 2, correctly generated code was defined as the total number of pre-annotated acronyms/abbreviations with the top code selected by Phase 2 annotator from Phase 1 annotations (the best). In the Relaxed Accuracy for Task 2, *correctly generated code* was defined as the total number of pre-annotated acronyms/abbreviations for which the code is contained in a list of possibly matching codes generated by the Phase 2 and 3 annotators (n-best).

In Task 3, the official primary and secondary measures were P@10 and NDCG@10 [40], respectively.[37] 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 (Fig. 1); NDCG@10 was computed using the original relevance assessments on a 4-point scale. The `trec_eval` evaluation tool[38] was used to calculate these evaluation measures[39]. Participants were also provided with other standard measures calculated by `trec_eval`[40].

The organisers provided the following evaluation tools on the Internet: an evaluation tool for calculation of the evaluation measures of Tasks 1a, 1b, and 2 as well as printing the results to a file; a Graphical User Interface (GUI) for calculation of the evaluation measures of Tasks 1a, 1b, and 2, as well as for

[37] Precision at 10 and Normalised Discounted Cumulative Gain at 10

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

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

[40] 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)

visualisation of system annotations against gold standard annotations; and a pointer to the `trec_eval` evaluation tool.

4 Results

The number of people who registered their interest in Tasks 1, 2, and 3 was 64, 56, and 55, respectively, and in total 34 teams with 18 unique affiliations submitted to the shared tasks (Table 1). No team participated in all three tasks. Teams represented China, France, India, Ireland, Republic of Korea, Spain, UK, 2 Australian states, and 8 US states. They had from 1 to 7 members (mean = 3.15, median = 3, and standard deviation = 1.52).

Teams submitted 113 systems (Table 2). 27 (7) were for Task 1a without (with) additional annotations. 21 (5) were for Task 1b without (with) additional annotations. 3 (2) were for Task 2 without (with) external annotations. 9 were participants' baseline systems for Task 3. In Task 3, 23 systems were not using the clinical reports nor additional annotations; 15 systems were using the clinical reports but without external annotations; and 1 system was using additional annotations but no clinical reports.

The number of teams that participated in Task 1a was 22. 5 of them were using additional annotations. 17 teams took the optional Task 1b. 4 of these teams were using additional annotations. 5 teams participated in Task 2, with 2 using additional annotations. 2 of the teams that participated in Task 2 also took Task 1a (but not Task 1b). 9 teams participated in Task 3 and only one of them was using additional annotations. All 9 participating teams submitted a baseline and systems not using the clinical reports nor additional annotations. 5 of the 9 teams also submitted systems using the clinical reports but without external annotations. 1 team submitted systems using external annotations but no clinical reports. 1 team participated in Tasks 2 and 3 and 1 team participated in Tasks 1a and 3, but these teams did not take any other tasks.

The best systems had an F1 score of 0.75 (0.80 Precision, 0.71 Recall) in Task 1a; Accuracies of 0.59 and 0.72 (0.66 without additional annotations) in Tasks 1b and 2; and P@10 of 0.52 in Task 3 (Tables 3–5). The use of additional annotations contributed to the system performance only in Task 2. In Task 3, the best system used the clinical reports. The best system that did not use the clinical reports came from the same team and had P@10 of 0.50.

The goal of the student mentoring track was to aid graduate students, regardless of which stage in their education they were in, and to provide additional feedback as a complement to their original advisors. This track was aimed at graduate students who would like to present and get more in-depth feedback on work related to the ShARe/CLEFeHealth2013 shared tasks or other relevant work in this research area, and included a peer-review process along with the assignment of one mentor (senior researcher) to provide constructive feedback in the CLEF conference on an extended abstract submission (2 pp.). The track received one submission.

Table 1. Participating teams. Some teams evolved during the shared tasks. For example, the Western Virginia University (WVU) had first six student-lecturer teams in Task 1 (i.e., WVU.AJ&VJ, WVU.AL&VJ, WVU.DG&VJ, WVU.FP&VJ, WVU.RK&VJ, and WVU.SS&VJ). Then the six students and their lecturer combined their forces for Task 2. Moreover, from the same organisation, many teams with changing team members participated (e.g., teams AEHRC and Mayo). Finally two teams with no members in common from the Seoul National University College of Medicine (i.e., MEDINFO and SNUBME) participated, knowing or not knowing about each other. In order to ease comparisons of organisations and countries, the organisers renamed the teams based on their affiliation (e.g., SNUBME.A for MEDINFO and SNUBME.B for SNUBME). This renaming was based on author names, affiliations, and team descriptions associated with the team submission.

ID	Organisers' team	Original team	Affiliation	Location	Number of participants
1	AEHRC.A	AEHRC	The Australian e-Health Research Centre, CSIRO, and Queensland University of Technology	QLD, Australia	3
2	AEHRC.B	AEHRC	"	"	3 (1 + 2 from A)
3	CLEAR	CLEAR	University of Colorado Boulder	CO, USA	1
4	CORAL	CORAL	The University of Alabama at Birmingham	AL, USA	3
5	HealthLanguageLABS	HealthLanguageLABS	Health Language Laboratories and The University of Sydney	NSW, Australia	3
6	KPSCMI	KPSCMI	Kaiser Permanente	CA, USA	3
7	LIMSI	LIMSI	LIMSI-CNRS	France	5
8	Mayo.A	Mayo	Mayo Clinic	MN, USA	4
9	Mayo.B	Mayo	"	"	5 (A + 1)
10	Mayo.C	Mayo	"	"	5 [3 + 2 from B (with 1 from A)]
11	NCBI	NCBI	National Center for Biotechnology Information, NLM/NIH/HHS	MD USA	3

ID	Organisers' team	Original team	Affiliation	Location	Number of participants
12	NIL-UCM	NIL-UCM	Universidad Complutense de Madrid	Spain	4
13	OHSU	ohsu	Oregon Health and Science University	OR, USA	3
14	QUT	QUT-TOPSIG	Queensland University of Technology	QLD, Australia	1
15	RelAgent	RelAgent	RelAgent Private Lt	India	2
16	SNUBME.A	SNUBME	Seoul National University College of Medicine	Republic of Korea	3
17	SNUBME.B	MEDINFO	"	"	2
18	THCIB.A	THCIB	Tsinghua University and Canon Information Technology (Beijing)	China	4
19	THCIB.B	THCIB	"	"	6 (A + 2)
20	UCDCSI.A	UCDCSI University College Dublin	Ireland	3	
21	UCDCSI.B	UCDCSI	"	"	2 (A − 1)
22	UCSC.CW&RA	UCDCSI	University of California, Santa Cruz	CA, USA	2
23	UCSC.KC&RA	KC	"	"	2
24	UOG	uogTr	Univeristy of Glasgow	UK	1
25	UTHealthCCB.A	UTHealth_CCB	The University of Texas Health Science Center at Houston	TX, USA	4
26	UTHealthCCB.B	UTHealth_CCB	"	"	5 (2 + 3 from A)
27	UTHealthCCB.C	UTHealth_CCB	"	"	6 (A + 2)
28	WVU	WVU	West Virginia University	VW, USA	7
29	WVU.AJ&VJ	ArvindWVU	"	"	2
30	WVU.AL&VJ	alamb	"	"	2
31	WVU.DG&VJ	Diganesan	"	"	2
32	WVU.FP&VJ	FAYOLA	"	"	2
33	WVU.RK&VJ	Rahul	"	"	2
34	WVU.SS&VJ	steven_seeger	"	"	2

Table 2. The tasks that the teams participated in. The suffix ".add" refers to using additional annotations. In Task 3, "*" indicates that clinical reports were used. The CORAL systems for Task 1b were not in the results announced on May 14 due to a missing registration until 17 Jun.

ID	Team	Number of submitted systems per task									
		1a	1a.add	1b	1b.add	2	2.add	3 baseline	3	3*	3.add
1	AEHRC.A	2		2							
2	AEHRC.B							1	3		
3	CLEAR	2		2							
4	CORAL	2		2							
5	HealthLanguageLABS	1				1					
6	KPSCMI	2		1							
7	LIMSI	2				1					
8	Mayo.A	1		2							
9	Mayo.B	1									
10	Mayo.C							1	3	3	
11	NCBI	2		2							
12	NIL-UCM	2		2							
13	OHSU							1	1		1
14	QUT							1	2	3	
15	RelAgent		2								
16	SNUBME.A	2									
17	SNUBME.B							1	3	3	
18	THCIB.A		1		1						
19	THCIB.B						1	1	3	3	
20	UCDCSI.A	2									
21	UCDCSI.B			2							
22	UCSC.CW&RA		2		2						
23	UCSC.KC&RA							1	2	3	
24	UOG							1	3		
25	UTHealthCCB.A	2		2							
26	UTHealthCCB.B					1					
27	UTHealthCCB.C							1	3		
28	WVU						1				
29	WVU.AJ&VJ	1		1							
30	WVU.AL&VJ		1		1						
31	WVU.DG&VJ	1		1							
32	WVU.FP&VJ	1		1							
33	WVU.RK&VJ		1		1						
34	WVU.SS&VJ	1		1							
	Systems:	27	7	21	5	3	2	9	23	15	1
	Teams:	17	5	13	4	3	2	9	9	5	1

Table 3. Evaluation in Task 1a. For the column of Strict F1 score, "*" indicates that the F1 score of the system was significantly better than the one immediately below (random shuffling, $p < 0.01$).

System ID ({team}.{system})	Strict Evaluation			Relaxed Evaluation		
	Precision	Recall	F1 score	Precision	Recall	F1 score
No additional annotations:						
(UTHealthCCB.A).2	0.800	0.706	0.750*	0.925	0.827	0.873
(UTHealthCCB.A).1	0.831	0.663	0.737*	0.954	0.774	0.854
NCBI.1	0.768	0.654	0.707*	0.910	0.796	0.849
NCBI.2	0.757	0.658	0.704*	0.904	0.805	0.852
CLEAR.2	0.764	0.624	0.687*	0.929	0.759	0.836
(Mayo.A).1	0.800	0.573	0.668*	0.936	0.680	0.787
(UCDCSI.A).1	0.745	0.587	0.656	0.922	0.758	0.832
CLEAR.1	0.755	0.573	0.651*	0.937	0.705	0.804
(Mayo.B).1	0.697	0.574	0.629*	0.939	0.766	0.844
CORAL.2	0.796	0.487	0.604	0.909	0.554	0.688
HealthLanguageLABS.1	0.686	0.539	0.604*	0.912	0.701	0.793
LIMSI.2	0.814	0.473	0.598*	0.964	0.563	0.711
LIMSI.1	0.805	0.466	0.590	0.962	0.560	0.708
(AEHRC.A).2	0.613	0.566	0.589*	0.886	0.785	0.833
(WVU.DG&VJ).1	0.614	0.505	0.554*	0.885	0.731	0.801
(WVU.SS&VJ).1	0.575	0.496	0.533	0.848	0.741	0.791
CORAL.1	0.584	0.446	0.505	0.942	0.601	0.734
NIL-UCM.2	0.617	0.426	0.504	0.809	0.558	0.660
KPSCMI.2	0.494	0.512	0.503*	0.680	0.687	0.684
NIL-UCM.1	0.621	0.416	0.498	0.812	0.543	0.651
KPSCMI.1	0.462	0.523	0.491*	0.651	0.712	0.680
(AEHRC.A).1	0.699	0.212	0.325*	0.903	0.275	0.422
(WVU.AJ&VJ).1	0.230	0.318	0.267*	0.788	0.814	0.801
UCDCSI.2	0.268	0.175	0.212*	0.512	0.339	0.408
SNUBME.2	0.191	0.137	0.160*	0.381	0.271	0.317
SNUBME.1	0.302	0.026	0.047	0.504	0.043	0.079
(WVU.FP&VJ).1	0.024	0.446	0.046	0.088	0.997	0.161
Additional annotations:						
(UCSC.CW&RA).2	0.732	0.621	0.672	0.883	0.742	0.806
(UCSC.CW&RA).1	0.730	0.615	0.668*	0.887	0.739	0.806
RelAgent.2	0.651	0.494	0.562*	0.901	0.686	0.779
RelAgent.1	0.649	0.450	0.532	0.913	0.636	0.750
(WVU.AL&VJ).1	0.492	0.558	0.523*	0.740	0.840	0.787
(THCIB.A).1	0.445	0.551	0.492*	0.720	0.713	0.716
(WVU.RK&VJ.1	0.397	0.465	0.428	0.717	0.814	0.762

Table 4. Evaluation in Tasks 1b and 2. For the column of Strict Accuracy, "*" indicates that the Accuracy of the system was significantly better than the one immediately below (random shuffling, $p < 0.01$).

System ID ({team}.{system})	Strict Accuracy	Relaxed Accuracy
Task 1b, no additional annotations:		
NCBI.2	0.589*	0.895
NCBI.1	0.587*	0.897
(Mayo.A).2	0.546*	0.860
(UTHealthCCB.A).1	0.514*	0.728
(UTHealthCCB.A).2	0.506	0.717
(Mayo.A).1	0.502*	0.870
KPSCMI.1	0.443*	0.865
CLEAR.2	0.440*	0.704
CORAL.2	0.439*	0.902
CORAL.1	0.410*	0.921
CLEAR.1	0.409*	0.713
NIL-UCM.2	0.362	0.850
NIL-UCM.1	0.362*	0.871
(AEHRC.A).2	0.313*	0.552
(WVU.SS&VJ).1	0.309	0.622
(UCDCSI.B).1	0.299*	0.509
(WVU.DG&VJ).1	0.241	0.477
(AEHRC.A).1	0.199*	0.939
(WVU.AJ&VJ).1	0.142	0.448
(WVU.FP&VJ).1	0.112*	0.252
(UCDCSI.B.2)	0.006	0.035
Task 1b, additional annotations:		
(UCSC.CW&RA).2	0.545*	0.878
(UCSC.CW&RA).1	0.540*	0.879
(THCIB.A).1	0.470*	0.853
(WVU.AL&VJ).1	0.349*	0.625
(WVU.RK&VJ).1	0.247	0.531
Task 2, no additional annotations:		
(UTHealthCCB.B).1	0.719*	0.725
(UTHealthCCB.B).2	0.683*	0.689
LIMSI.1	0.664*	0.672
TeamHealthLanguageLABS.1	0.467	0.488
Task 2, additional annotations:		
(THCIB.B).1	0.657*	0.685
WVU.1	0.426	0.448

Table 5. Evaluation in Task 3. Result which are significantly worse than the baseline for P@10 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 BM25 retrieval model and relevance feedback (BM25_FB). The format of Run ID ({team}.{run}.{rank}) is defined in Section 3.3. The best P@10 values for each team is *emphasised*.

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

5 Discussion

This paper reported on a novel evaluation lab with an aim to support the continuum of care by developing methods and resources that make clinical reports in English easier to understand for patients. This ShARe/CLEFeHealth2013 lab had a mentoring track for graduate students and three shared tasks: identification and normalisation of disorders in clinical reports with respect to terminology standards in healthcare; normalisation of abbreviations and acronyms in clinical reports with respect to terminology standards in healthcare; and IR to address questions patients may have when reading clinical reports. The focus on patients' information needs as opposed to the specialised information needs healthcare workers was the main distinguishing feature of the lab from previous shared tasks on NLP and ML. The lab attracted a substantial amount of interest and demonstrated the capabilities of submitted systems and participating teams in making clinical reports easier to understand for patients. Over 30 teams from America, Asia, Australia, and Europe submitted altogether 113 systems to the shared tasks. The best systems had the F1 score of 0.75 in Task 1a; Accuracies of 0.59 and 0.72 in Tasks 1b and 2; and Precision at 10 of 0.52 in Task 3.

The significance of the lab was emphasised by the organisers' making the text documents, annotations, queries, mappings between queries and the matching clinical report, the matching result sets, relevance assessments, and evaluation tools available for future research and development. The lab developed new annotated datasets, including English text from clinical reports and the Internet. De-identified clinical reports for Task 1–3 were from US intensive care and Task 3 also used documents from the Internet. Task 1 annotations originated from the ShARe annotations, but for Tasks 2 and 3, new annotations, queries, and relevance assessments were created. Guidelines[41] for human subjects training, ethics clearance, research permission, registration, user access, data/annotation format, tools, and contact people were made available.

These three tasks have all aimed at supporting the patient, potential patient or next of kin to understand and have a better picture of their health condition. By working towards easier-to-understand translations of clinical text, we support the patient empowerment and patients' ability to make informed decisions concerning their own health and care.

Acknowledgement

The ShARe/CLEFeHealth2013 evaluation lab has been supported in part by (in alphabetical order) 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 CLEF Initiative; the European Science Foundation (ESF) project ELIAS, Evaluating Information Access Systems; the Khresmoi project, funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 257528; the ShARe project

[41] `https://sites.google.com/site/shareclefehealth/`

funded by the United States National Institutes of Health (R01GM090187); the US Department of Veterans Affairs (VA) Consortium for Healthcare Informatics Research (CHIR); the US Office of the National Coordinator of Healthcare Technology, Strategic Health IT Advanced Research Projects (SHARP) 90TR0002; and the Vårdal Foundation (Sweden).

We acknowledge the generous support of time and expertise that the annotators as well as members of the organising and mentoring committees have invested in this first pilot year of the evaluation lab. We want to thank the following individuals: Allan Hanbury (Vienna University of Technology, Austria); Anni Järvelin and Dimitrios Kokkinakis (University of Gothenburg, Sweden); Digvijay Khangarot, Thomas Souchen, Timothy Sladden, and Warren Brown (Australian E-Health Research Centre, CSIRO, QLD, Australia); Erika Siirala, Filip Ginter, Heljä Lundgren-Laine, Jenni Lahdenmaa, Laura Maria Murtola, Lotta Kauhanen, Marita Ritmala-Castren, Riitta Danielsson-Ojala, Saija Heikkinen, and Sini Koivula (University of Turku, Finland); Jussi Karlgren (Gavagai and KTH Royal Institute of Technology, Sweden); Hans Moen (Norwegian University of Science and Technology, Norway); Henning Müller (University of Applied Sciences Western Switzerland); Hercules Dalianis (DSV Stockholm University, Sweden); Maricel Angel (NICTA, ACT, Australia); Özlem Uzuner (State University of New York, NY, USA); Pamela Forner (CELCT Center for the Evaluation of Language and Communication Technologies and CLEF, Trento, Italy); Preben Hansen (Stockholm University, Sweden); Qing Treitler Zeng and Tyler Forbush (University of Utah, SLC VA, UT, USA); and Rune Saetre (Norwegian University of Science and Technology, Norway).

References

1. Allvin, H., Carlsson, E., Dalianis, H., Danielsson-Ojala, R., Daudaravicius, V., Hassel, M., Kokkinakis, D., Lundgren-Laine, H., Nilsson, G., Nytro, O., Salanterä, S., Skeppstedt, M., Suominen, H., Velupillai, S.: Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. Journal of Biomedical Semantics **2**(Suppl 3) (2011) S1
2. 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)
3. Fox, S.: Health Topics: 80% of internet users look for health information online. Technical report, Pew Research Center (February 2011)
4. Kummervold, P., Chronaki, C., Lausen, B., Prokosch, H., Rasmussen, J., Santana, S., Staniszewski, A., Wangberg, S.: eHealth trends in Europe 2005–2007: A population-based survey. Journal of Medical Internet Research **10**(4) (2008) e42
5. Experian Hitwise: Google Receives 87.81 Percent of Australian Searches in June 2008. In: `http://www.hitwise.com/au/press-centre/press-releases/2008/ap-google-searches-for-june/`. (2008)
6. 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)
7. 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)

8. 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)
9. Becker, H.: Computerization of patho-histological findings in natural language. Pathologia Europaea **7**(2) (1972) 193–200
10. Anderson, B., Bross, I., Sager, N.: Grammatical compression in notes and records: Analysis and computation. American Journal of Computational Linguistics **2**(4) (1975) 68–82
11. Hirschman, L., Grishman, R., Sager, N.: From text to structured information: automatic processing of medical reports. In: American Federation of Information Processing Societies: 1976 National Computer Conference. Volume 45 of AFIPS Conference Proceedings., New York, NY, USA, Association for Computational Linguistics (1976) 267–275
12. Collen, M.: Patient data acquisition. Medical Instrumentation **12**(Jul–Aug) (1978) 222–225
13. Sarkar, I.: Biomedical informatics and translational medicine. Review. Journal of Translational Medicine **8** (2010) 22
14. Demner-Fushman, D., Chapman, W., McDonald, C.: What can natural language processing do for clinical decision support? Review. Journal of Biomedical Informatics **42**(5) (2009) 760–772
15. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.: Extracting information from textual documents in the electronic health record: a review of recent research. Review. Yearbook of Medical Informatics (2008) 128–44
16. Reiner, B., Knight, N., Siegel, E.: Radiology reporting, past, present, and future: the radiologist's perspective. Review. Journal of the American College of Radiology: JACR **4**(5) (2007) 313–319
17. Suominen, H., Lehtikunnas, T., Back, B., Karsten, H., Salakoski, T., Salanterä, S.: Applying language technology to nursing documents: pros and cons with a focus on ethics. Review. International Journal of Medical Informatics **76**(Suppl 2) (2007) S293–S301
18. Zweigenbaum, P., Demner-Fushman, D., Yu, H., Cohen, K.: Frontiers of biomedical text mining: current progress. Review. Briefings in Bioinformatics **8**(5) (2007) 358–375
19. Mendonça, E., Haas, J., Shagina, L., Larson, E., Friedman, C.: Extracting information on pneumonia in infants using natural language processing of radiology reports. Journal of Biomedical Informatics **38**(4) (2005) 314–321
20. Pakhomov, S., Buntrock, J., Chute, C.: Automating the assignment of diagnosis codes to patient encounters using example based and machine learning techniques. Journal of the American Medical Informatics Association: JAMIA **13**(5) (2006) 516–525
21. Chapman, W., Nadkarni, P., Hirschman, L., D'Avolio, L., Savova, G., Uzuner, Ö.: Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. Editorial. Journal of the American Medical Informatics Association : JAMIA **18** (2011) 540–543
22. Robertson, S., Hull, D.: The TREC-9 filtering track final report. In: NIST Special Publication 500-249: The 9th Text REtrieval Conference (TREC 9). (2000) 25–40
23. Roberts, P.M., Cohen, A.M., Hersh, W.R.: Tasks, topics and relevance judging for the TREC genomics track: five years of experience evaluating biomedical text information retrieval systems. Information Retrieval **12** (2009) 81–97

24. Voorhees, E.M., Tong, R.M.: Overview of the TREC 2011 medical records track. In: Proceedings of TREC, NIST (2011)
25. Kalpathy-Cramer, J., Müller, H., Bedrick, S., Eggel, I., de Herrera, A., Tsikrika, T.: The CLEF 2011 medical image retrieval and classification tasks. In: Working Notes of CLEF 2011 (Cross Language Evaluation Forum). (2011)
26. Müller, H., Clough, P., Deselaers, T., Caputo, B., eds.: Experimental Evaluation in Visual Information Retrieval. Volume 32 of The Information Retrieval Series. Springer (2010)
27. Uzuner, Ö., South, B., Shen, S., DuVall, S.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association : JAMIA **18** (2011) 552–556
28. Pestian, J., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K., Duch, W.: A shared task involving multi-label classification of clinical free text. In: BioNLP Workshop of the Association for Computational Linguistics, Association for Computational Linguistics (2007) 97–104
29. Pestian, J., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, Ö., Wiebe, J., Cohen, K., Hurdle, J., Brew, C.: Sentiment analysis of suicide notes: A shared task. Biomedical Informatics Insights **5**(Suppl 1) (2012) 3–16
30. 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, Khresmoi EU project (2012)
31. Hanbury, A., Müller, H.: Khresmoi – multimodal multilingual medical information search. In: MIE village of the future. (2012)
32. Bodenreider, O., McCray, A.: Exploring semantic groups through visual approaches. Journal of Biomedical Informatics **36** (2003) 414–432
33. South, B.R., Shen, S., Leng, J., Forbush, T.B., DuVall, S.L., Chapman, W.W.: A prototype tool set to support machine-assisted annotation. In: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. BioNLP '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 130–139
34. 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)
35. Koopman, B., Zuccon, G.: Relevation! an open source system for information retrieval relevance assessment. arXiv preprint (2013)
36. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11) (1975) 613–620
37. Robertson, S.E., Jones, S.: Simple, proven approaches to text retrieval. Technical Report 356, University of Cambridge (1994)
38. Yeh, A.: More accurate tests for the statistical significance of result differences. In: Proceedings of the 18th Conference on Computational Linguistics (COLING), Saarbrücken, Germany (2000) 947–953
39. 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, New York, NY, USA, Association for Computing Machinery (2007) 623–632
40. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4) (2002) 422–446