Navigating through Very Large Sets of Medical Records: An Information Retrieval Evaluation Architecture for Non-standardized Text

Markus Kreuzthaler, Marcus Bloice,
Klaus-Martin Simonic, and Andreas Holzinger

Institute for Medical Informatics, Statistics and Documentation,
Auenbruggerplatz 2, 8036 Graz, Austria
{markus.kreuzthaler,marcus.bloice,klaus.simonic,
andreas.holzinger}@medunigraz.at

Abstract. Despite the prevalence of informatics and advanced information systems, there exists large amounts of unstructured text data. This is especially true in medicine and health care, where free text is an indispensable part of information representation. In this paper, the motivation behind developing information retrieval systems in medicine and health care is described. An overview of information retrieval evaluation is given, before describing the architecture and the development of an extendible information retrieval evaluation framework. This framework allows different information retrieval tools to be compared to a gold standard in order to test its effectiveness. The paper also gives a review of available gold standards which can be used for research purposes in the area of information retrieval of medical free texts.

Keywords: information retrieval, medicine, text mining, evaluation, gold standards.

1 Introduction and Motivation

Retrieving information from a large amount of medical free text is an important factor in clinical research, quality assurance, and medical accounting [1]. Also, with the introduction of large medical information systems, legacy texts are being processed by optical character recognition systems, digitized, and fed into archive systems. Medical professionals are often confronted with masses of free text and must deal with the task of finding the relevant information. Consequently, there is a need for smart information retrieval systems that can operate on this type of text, and the current systems in use are far from being infallible [2]. We believe that to make this data more accessible, usable, and useful, both the technological aspects, as well as human aspects, must be taken into consideration [3].

Although language understanding in general is still an unsolved problem, restricted domains such as medical notes and letters seem to be more tractable. This holds even more for the task of information retrieval where a thorough text understanding is not required and the identification of the concepts mentioned

A. Holzinger and K.-M. Simonic (Eds.): USAB 2011, LNCS 7058, pp. 455–470, 2011.

in a text together with the detection of their polarity suffices in most cases. However, the medical language used in our system poses a number of particular challenges:

- Abbreviations are frequent, often ambiguous or even ad hoc.
- There may be typos and optical character recognition (OCR) errors.
- A mixture of different languages (e.g. German, Latin, English) is used together with expressions of medical jargon.
- The same concept may be expressed in various ways using synonyms and linguistic variations.

As far as the context is concerned, homonyms are often only resolvable when knowing the context in which they belong. For a doctor, as long as the free text content is clear, correct spelling is not entirely important.

Meeting the information demand of medical professionals poses challenges that go beyond classical information retrieval. The first is robustness with respect to the mentioned *variability* and *noise* of the underlying texts. Another challenge is to take into account the semantic relations in the medical domain. These involve, in particular, taxonomic, functional, and anatomic relations. For instance, searching for reports of neoplasms in the gastrointestinal tract means having to match to all kinds of tumors and carcinomas in many parts of the body, such as the esophagus, stomach, intestines etc.

Of course, it must also be possible to evaluate a system's performance, in order to compare it with other solutions currently being used or systems that are under development. Furthermore, it is of importance to be able to compare any system's performance against a human expert who is executing the same task. Human experts possess knowledge about the typical idiosyncrasies that are contained within free texts, and are able to apply previous experience and knowledge to the search request to get high precision *and* high recall values from the information retrieval process. On the other hand they must use much longer expressions in their chosen query language to get the required information, which contrasts the way semantic information retrieval works; the input to the tool is in a human readable form.

This paper, besides providing a review of information retrieval testing and available gold standards that can be used for this purpose, concentrates on the description of a test environment for information retrieval systems working on medical free text copora. The test environment allows the comparison between the human expert's results and the retrieval tool's results according to an information need. The core elements of the evaluation framework that must exist were identified, and these components were grouped together as one graphical user interface (GUI) in order to provide a user-friendly front end to the system, following usability engineering methods [4]. The GUI supports the speedy evaluation of an information retrieval tool (IR Tool) and also aids the developer of the tool in further enhancing it.

The remainder of the paper is organized as follows: Section 2 motivates information retrieval in medicine and health care and gives a review regarding information retrieval evaluation. Furthermore, openly available gold standards

that can be used for free text information retrieval testing are listed. Section 3 describes the entire evaluation architecture identifying the main components that the architecture consists of and logically groups them together. Section 4 depicts the results of the evaluation environment concentrating on the arrangement of components in the GUI. Finally, Section 5 concludes the paper and gives an outlook of open questions and future research directions in this area.

2 Theoretical Background and Related Work

In this section we are motivating information retrieval in the context of medical textual data. Afterwards, a short overview of information retrieval testing is given, concentrating on available gold standards which can be used for information retrieval research. This is an important aspect due to the fact that in order to test any information retrieval system's usefulness in the medical domain, data must be used which, by its nature, is *very sensitive*.

2.1 Information Retrieval in Medicine and Health Care

To enhance quality of patient care and to provide a better use of evidence, information retrieval systems are increasingly used by physicians [5]. Information retrieval system evaluation is an ongoing are of research [6,7,8] and advanced text mining techniques help to handle the information overload, for example in medical literature [8,9]. In contrast to this, text mining methods lack enhancement in the area of medical information and documentation systems [1,3,10]. Sophisticated medical information retrieval systems have to handle very large sets of medical documents, where typically *non-standardized* text makes up a significant amount of this patient data. This data, often called free text in literature, has been very long in the focus of research [11,12,13] and has not lost its importance [14]. The automatic analysis of text is still a challenging problem [15,16,17], which is in contrast to the effort which has to be made to produce text. Relevant relationships can stay completely undiscovered, because relevant data are scattered and no investigator has linked them together manually [18].

2.2 Information Retrieval Evaluation

According to [19] the standard procedure to measure information retrieval effectiveness comprises; a document collection, a test suite of information requests expressible as queries and a set of judgments for each query-document pair, which defines each pair as either relevant or not relevant. The result of this binary classification process to an information need is called a *ground truth judgment of relevance* or simply a *gold standard*. Another important fact is that the test collection should be a sample of the kinds of text that will be encountered in the operational setting of interest and particularly for testing information retrieval systems that work on medical free text data, it is hard to find utilizable corpora that can be used for information retrieval research [20]. In extension

to [20] we want give a review of openly available free text corpora which can be used as gold standards for research purposes (Table 1). We therefore concentrate our review on copora containing text types which pose the challenges mentioned in Section 1 which differ significantly from text corpora used in the field of biomedical natural language processing (bioNLP) for example. Typically in bioNLP most of the annotated texts are drawn from biology literature or from abstracts of scientific literature (e.g. MEDLINE), so *not* reflecting the difficulties in computer based patient records. For the sake of completeness we want to mention The National Centre for Text Mining (NaCTeM)[1] which is an excellent resource finding corpora which can be used for bioNLP. Table 1 which is an extension to [21] gives an overview of available test corpora for medical free text information retrieval and extraction evaluation. An excellent overview to other copora which can be used for natural language processing (NLP) can be found at `http://nlp.stanford.edu/links/statnlp.html#Corpora`.

Table 1. Available free text gold standards in the medical domain

Corpus Name	*Description*
Computational Medicine Challenge	1954 reports are assigned with ICD-9-CM codes from a radiology department. The corpus was initially used for testing and evaluating different classifiers [22].
ImageCLEFMed	The data set comprises ca. 50 000 images with textual annotations. Even though made up for content based image retrieval, the textual content to each picture plus annotations can be used as a gold standard for text based information retrieval testing [23,24].
Ogren	The corpus contains 60 clinical notes annotated with functional disorders. [25].
CLEF Corpus	The set exits of 167 medical documents annotated with diseases, drugs, body regions etc. as well as their internal relations [26].
i2b2	The text corpus comprises 889 anonymized epicrises and contains annotation which can be used for evaluating algorithms for the task of de-identification of clinical records. A subset of the texts further on contains annotations about the smoking status of the patient [27,28].
BLULab	The BLULab NLP Repository contains a collection of ca. 100 000 de-identified clinical records from multiple U.S. hospitals during 2007. The reports are annotated with ICD-9 Codes. `http://nlp.dbmi.pitt.edu/nlprepository.html`

The Text REtrieval Conference (TREC, `http://trec.nist.gov/tracks.html`) has recently started a track with the following aim: "The goal of the Medical Records track is to foster research on providing content-based access to the free-text fields of electronic medical records." The BLULab corpus will

[1] `http://www.nactem.ac.uk/`

be used for this purpose and judged information needs through pooling will be available soon.

Having chosen a collection of interest, a metric must be decided upon. Typically one differs between metrics for *ranked* (Precision, Recall, Fallout, F-Measure) and *unranked retrieval results* (R-Precision, Precision at k, Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NCDG)) [29,30]. Beside pure statistical evaluation metrics other levels of testing should be considered. [31] stated testing at the engineering level, the input level, the processing level, the output level, the use and user level, and the social level that should be accounted for. Furthermore the human factor and human information behavior [32], in the context of information retrieval systems, are important factors when developing such systems. The systems must be capable of satisfying user needs and therefore human factors play an important role [33] for the acceptance of the retrieval systems as a whole.

3 Methods and Materials

This section describes the generic architecture that was developed for information retrieval testing based on *medical* free text corpora. We start by specifying the objective target and requirements of the overall system. Based on these requirements main components are identified, which are described in Section 3.2. The next section concentrates on the human computer interface (HCI) aspects that were considered when developing the GUI for the test framework, by logically grouping together the mentioned components. At the end we depict the overall architecture, containing the different parts identified previously.

3.1 Objective Target and Requirements

The main objective was to evaluate an information retrieval system within an integrated test environment and to compare the performance achieved by the IR Tool to that of a human expert who searched for the same task. As a representation metric the precision, recall, fallout, and F-measure operating figures were chosen to gauge the performance of the IR Tool. A further requirement for the test environment was that different versions of one IR Tool and their corresponding evaluation values could be saved. As a consequence, different versions can be loaded and compared according to their performance values. Besides testing different versions of one IR Tool, the framework working in the background also supports the integration and test of different IR Tools. Therefore the main requirements for the system are:

- To support the developer in enhancing the IR Tool. Therefore, both a test set (3034 ICD-annotated pathology reports) and a training set (508 ICD-annotated pathology reports) were created. The test set allows the developers to only see the statistical evaluation. The training set, on the other hand, provides the developer with the chance to view details as to why a text was found or not found by the IR Tool.

- Evaluation results with preset information retrieval parameters can be saved and compared to evaluation results from previous test runs. For example two runs with different information retrieval parameters.
- A user friendly graphical interface, which logically groups together components in a consistent way. In particular, statistical evaluation values between different IR Tool versions, and between the training and test data, were arranged in a way that makes them easy to compare them at a first glance.
- To be easily extendable to accommodate new test or training sets of different text types. For example, pathology inflammation and radiology thorax.
- The framework should also support the integration of different IR Tools. This makes it possible not just to compare different versions of one tool but different versions of different IR Tools.
- To make it easy to navigate through the different result sets of the IR Tool (true positives, false positives, false negatives, true negatives and all found).
- To highlight texts that differ between different result sets of the expert and the IR Tool according to an information need, to gain quick access.
- To ensure that the evaluation environment is accessible online with different user rights for different developers.

3.2 Component Identification

Considering the requirements from the previous sections, we identified the core components that the evaluation architecture must consist of. These components are:

Data Base. A data base specifies a certain type of medical free text reports. The text's type in our case are pathology reports. It is the data that the IR Tool must perform actions on.

Training Data. The training data is an annotated set of medical reports which belong to a certain data base. All medical reports which belong to a certain training data type can be fully accessed, so all details belonging to this text are displayed. This means that it is possible to enhance the information retrieval system for this particular type of data.

Test Data. The test data is an annotated set of medical reports which belong to a certain data base. No medical reports that belong to a certain test data type can be accessed, and therefore no details belonging to this text are displayed.

Test Case. A test case defines a specific set of queries. For example, the test case "One Word" comprised the queries "Hepatitis", "Appendicitis", "Colitis", and "Gastritis".

Query. A query belongs to one or many test cases. A query on its own consists of three search requests.

Search Request. There are three types of search requests.

- *Expert Search Request.* This is the human expert's search statement used to fulfill the information retrieval task.

- *Information Retrieval Search Request.* This is the statement that is used as an input for the IR Tool.
- *Gold Standard Search Request.* This is the statement that is used to get the truth out of the gold standard. That means that reports that correctly belong to the diagnosis 'hepatitis' are returned, for example.

Test Case Result. An evaluated test case returns a test case result. A test case result consists of one or many query results.

Query Result. An evaluated query returns a query result. The query result also contains the results of different evaluation metrics.

Information Retrieval Tool. This is the IR Tool, where benchmarks should be performed on.

Information Retrieval Parameters. The information retrieval can be enhanced by further tuning specific parameters which influence the results.

Evaluation Metric. To determine the quality of the information retrieval system an adequate metric has to be used.

3.3 Component Grouping

Concerning the requirement that a user-optimized GUI had to be created, this section describes the logical grouping of the components identified in the previous section which are finally associated to a view. This grouping resulted in the following three views and functionalities:

Search and Evaluate. Abstractly spoken, a test case is used in conjunction with specific information retrieval parameters on training data and/or test data that belong to a certain data base, with the aim of evaluating the performance of the IR Tool. Therefore, the data, test case, query, and information retrieval parameter selection process should be grouped together in their own section of the GUI. We refer to these elements as *Komponenten Pool I.*

The evaluation of a test case produces a test case result. To cater for this, there must be a section in the GUI where different views of the test case result can be selected. This result view selection comprises of:

- *Data Selection.* This is used to choose whether to show the evaluation results that belong to the training data or to the test data.
- *Query.* This is used to select a specific query result from the test case result.
- *Sub Set.* This is used to select the amount of true positives, false positives, true negatives, false negatives or all found.

We refer to these elements as *Komponenten Pool II.*

The result view is the third section of the GUI. It comprises of the following components:

- Training data based evaluation results from the expert and the IR Tool.
- Test data based evaluation results from the expert and the IR Tool.

- A section that comprises of a list of found reports along with their identifiers, how many were found, and the search request that lead to the result. Such a section has to be set up for the Expert, the IR Tool and the Gold Standard.
- A field where query results can be annotated with further information.
- A field where any save options can be selected. This is required as it is possible for test case results to be saved to a data medium.

We refer to these elements as *Komponenten Pool III.*

Load and Compare. As mentioned in Section 3.1, it is essential that evaluation values for previous versions of the information retrieval system can be loaded and compared. To achieve this, the following elements are required:

- *Test Case Result Filter.* Because many different IR Tool versions will be tested using different test cases, it was deemed necessary to allow certain test case results to be filtered that match certain criteria.
- *Filtered Test Case Results.* If the filtering process results in different test case results, the user can choose to compare these.

Again we refer to these elements as *Komponenten Pool I.*

The comparison of two test case results should be feasible in a user friendly, well-arranged way. The following menu items are necessary:

- *Data Selection I.* This menu item allows one to toggle whether the results belonging to the training- or test data should be shown. The selection is connected to the first loaded test case result.
- *Query.* This menu item allows one to toggle which query result of the test case result should be shown. The selection is connected to both loaded test case results.
- *Sub Set.* This menu item allows one to toggle what result set should be shown (true positives, false positives, false negatives, true negatives and all found). The selection is connected to both loaded test case results.
- *Data Selection II.* This menu item allows one to toggle whether the results belonging to the training- or test data should be shown. The selection is connected to the second loaded test case result.

We refer to these elements as *Komponenten Pool II.*

Two loaded test case results must be compared in a user-friendly way with the following content:

- Evaluation results based on the selected data (either training data or test data) from the human expert and the IR Tool.
- A section that comprises of a list of found reports along with their identifiers, how many have been found, and what the search request was that led to the result.
- A field where query results can be annotated.

Both these fields must appear twice, so that it is possible for two case results to be compared. Again, we refer to these elements as *Komponenten Pool III.*

Details. When developing the IR Tool, it is possible to not only get evaluation values but also to get detailed information as to why a certain report was found or not. Developers of the IR Tool are allowed to have a certain amount of insight into these details, but only for reports which belong to the training data. If required, they can get detailed feedback regarding each report in order for them to be able to enhance the retrieval performance. Conversely, developers have no insight into the details of the medical text corpora that belong to the test data. Performing a test run using test data only returns evaluation values according to the chosen metrics, as stated in Section 3.2. Any further access is disallowed. Aside from the aforementioned details, developers are also able to change the inspected report in this view in order to inspect how the information retrieval system deals with the altered text. Developers can then see if the changed text is found or not, according the details contained within that text. With the IR Tool Console the following functionality and content is mapped:

- The full content of the report is displayed.
- The IR Tool parameters which are responsible for this result are displayed and can be altered.
- Information about the scoring-process regarding the reports affiliation to the information need can be read.
- The report can be altered and the outcome of the retrieval process using the altered text can be checked.
- The semantic presentation of the report (Concept Graph) is displayed.

These elements are referred to as *Komponenten Pool Details*.

3.4 Architecture

This section describes the architecture and the components within this architecture as displayed in Fig. 1. The design is a typical three-tier architecture, which is specified in more detail in the next paragraph.

Client Tier. Contains the views as described in Section 3.3. The user is interacting with the environment via a web-browser.

Logic Tier. The communication with the server is realized with the Spring MVC paradigm. Important server side components are the corresponding controllers for the views (Search and Evaluate, Load and Compare, Details). The controllers are using the Java-based evaluation framework to handle the test logic. The framework consists of an evaluation logic, communicating with three main components: Data Bank Handler (DB Handler), the Metrics Module (Metrics) containing different statistical evaluation measures and IR Tool handler (IR Handler).

Data Tier. The data tier exists of three elements:

- Every data basis is stored in a data bank, containing medical reports split up into training- and test data. Further on test cases, queries and corresponding results are saved. A new data basis (e.g. radiology thorax) is saved in a new data bank.

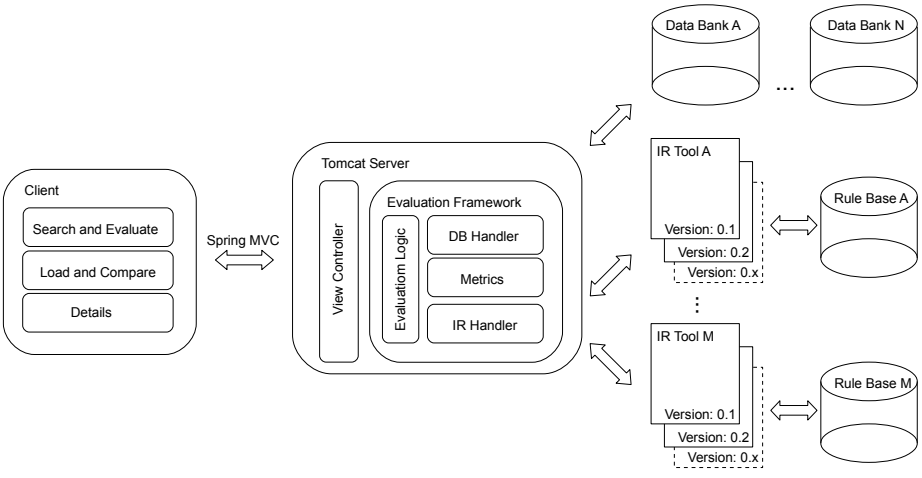

Fig. 1. Architecture of the evaluation environment

- The IR Tool under test. As well as the testing of different versions of *one* IR Tool is supported, *different* information retrievals tools can be integrated.
- Lots of information retrieval technologies communicate in the background with a rule base, which contains further logic that is needed for the retrieval process.

4 Results

This section describes the results of the GUI and provides some screen shots of the completed system. Therefore this section concentrates on the arrangement of the components of the developed information retrieval test environment. Statistical evaluation results and a detailed survey of the benchmark can be found in [34].

4.1 Hardware and Software Setup

The entire system was developed as an online, web application. The website itself runs on the Apache Tomcat web server, has a MySQL backend and was developed using the Spring Model-View-Controller paradigm in the Java programming language. The physical server runs the FreeBSD operating system on an Intel processor. The IR Tool under test was developed by ID Berlin (ID Information und Dokumentation im Gesundheitswesen GmbH & Co. KGaA ID) and uses the ID MACS® Server as rule base in the background. The IR Tool's retrieval process is based on a semantic text representation of the medical reports. This is achieved by applying a natural language processing technique to the texts, and the resulting concepts are parsed to a Wingert nomenclature [35,36,37].

4.2 Graphical User Interface

In Fig. 2 it is possible to see the arrangement of all important GUI components that were defined in Search and Evaluate. At the top of the web page all elements belonging to *Komponenten Pool I* are visible. Beneath these are the result view selection elements grouped into *Komponenten Pool II* - most of the page is dedicated to *Komponenten Pool III*. As mentioned in Section 3.1, one of the requirements was to compare the results achieved by an expert with the results achieved by the IR Tool's information retrieval process. Therefore, any different document identifiers are highlighted so that the user can see them at a glance. The evaluation results from the training data are placed next to the results from the test data, so that the user can quickly and directly compare the values achieved by the precision, recall, fallout, and F-measure algorithms. By clicking on any report, the user accesses the details of the text, assuming they have the sufficient access rights (Fig. 4).

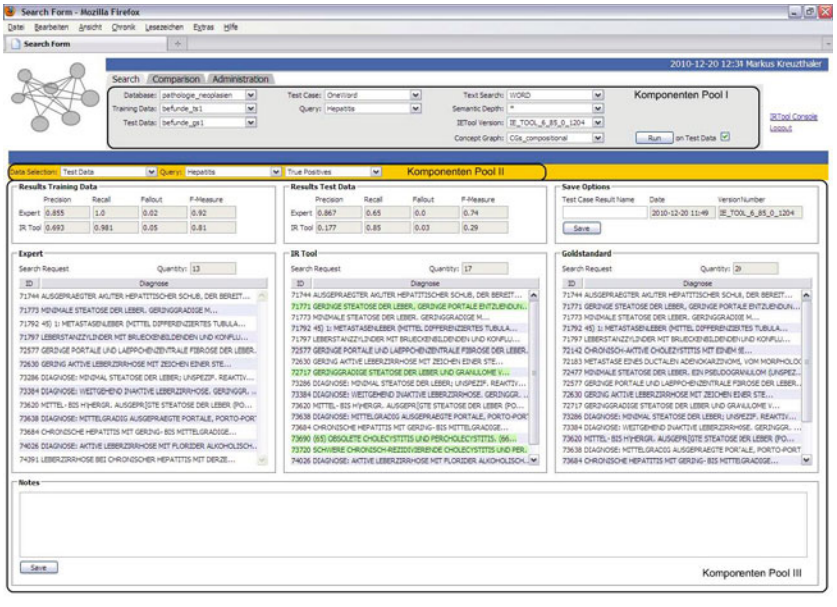

Fig. 2. Search and Evaluate View

The arrangement of the components we defined for the Load and Compare view is described in Fig. 3. Again, elements belonging to *Komponenten Pool I* are at the top of the screen where the user can select which test case results should be loaded. Beneath it is the result view section *Komponenten Pool II*, as well as *Komponenten Pool III*. As can been seen from Fig. 3, both loaded test case results are arranged in such a way as to allow the user to compare the different IR Tools easily.

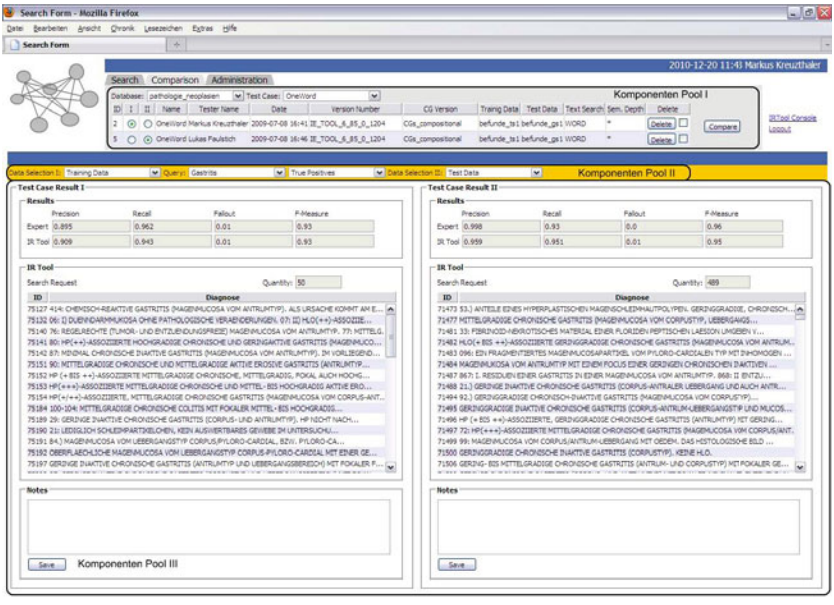

Fig. 3. Load and Compare View

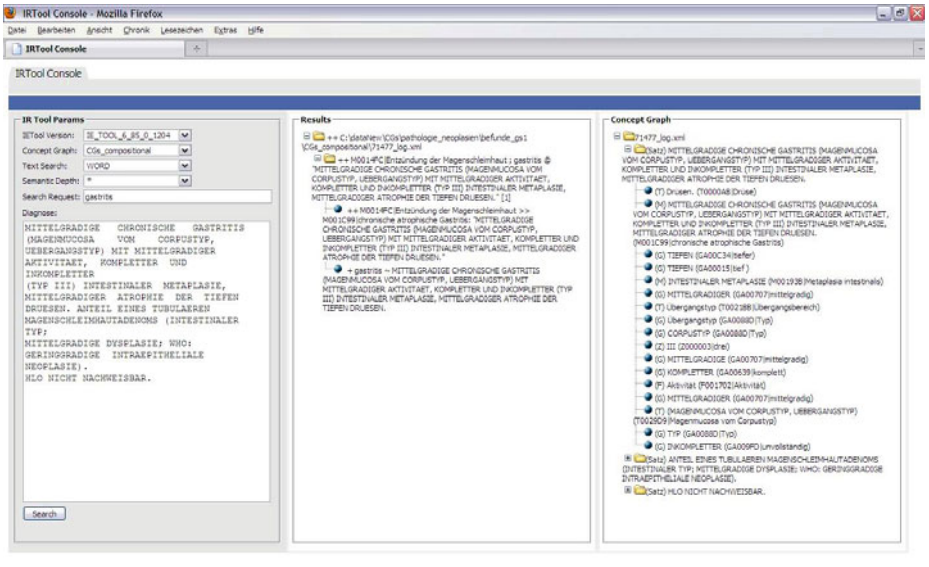

Fig. 4. Details View

By clicking on a report, the details view appears (Fig. 4). As can be seen from the image, the left side of the view shows the full report and all the IR Tool parameters. In the center of the view is the information regarding why the

diagnosis was found or not. On the right hand side, the results of the report's natural language processing and its mapping to the Wingert nomenclature is shown. Further, you can directly alter the text and repeat the search on the updated text by clicking on "Search".

5 Conclusion and Outlook

In this paper, do our best knowledge we tried to expand the review on available medical free text gold standards which was made by [21]. In addition, we presented an information retrieval evaluation architecture, described its core components, and explained the development of a user friendly GUI that groups these components in a logical and consistent way.

Future work will concentrate on how to dynamically integrate other information retrieval tools so that they can be compared and contrasted to one another. In the German speaking community there are just a few companies that provide solutions or prototypes for medical free text retrieval or information extraction (e.g. SemFinder, ID Berlin, Averbis). Instantiating a challenge using our evaluation architecture would be of interest.

The development of new gold standards to address other fields of medicine is also of the utmost importance. Furthermore, there are some tools that claim to be applicable to general free text medical corpora, and these should be rigorously tested. The kernel of our system, namely the gold standard developed specifically for this project's needs, currently contains only German texts. Multi-lingual information retrieval in the medical domain is the subject of ongoing research.

During the development of the tool, it became apparent that there is no open source ground truth for German medical free texts in existence [20]. The development of such open source ground truths for medical reports, which span several medical fields, and are multi-lingual, are of special interest and warrant much effort in the future. It is clear that if such open source gold standards were to exist, it would certainly help the information retrieval tool community and would definitely encourage others to join this field.

To conclude, information retrieval in the field of medicine is becoming increasingly important and relied upon. This is especially reflected through the fact that a Medical Records Track has been started recently at the TREC so the information retrieval community will become aware of the challenges medical free text retrieval pose. We previously stated that this topic should be introduced as a track to the TREC in [34]. Semantic search seems the only possible way to search medical free text corpora and return meaningful results, while at the same time this area of research is being hampered by a number of factors, such as a lack of an open source gold standard initiative in the German speaking community. The evaluation environment presented here constitutes our contribution towards medical free text retrieval.

References

1. Holzinger, A., Geierhofer, R., Errath, M.: Semantic information in medical information systems-from data and information to knowledge: Facing information overload. In: Proc. of I-MEDIA, vol. 7, pp. 323–330 (2007)
2. Buckley, C.: Why current ir engines fail. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 584–585. ACM (2004)
3. Holzinger, A., Geierhofer, R., Mödritscher, F., Tatzl, R.: Semantic information in medical information systems: Utilization of text mining techniques to analyze medical diagnoses. Journal of Universal Computer Science 14(22), 3781–3795 (2008)
4. Holzinger, A.: Usability engineering methods for software developers. Communications of the ACM 48(1), 71–74 (2005)
5. Hersh, W.R., Hickam, D.H.: How well do physicians use electronic information retrieval systems? JAMA: The Journal of the American Medical Association 280(15), 1347 (1998)
6. Robertson, S.E., Hancock-Beaulieu, M.M.: On the evaluation of ir systems. Information Processing & Management 28(4), 457–466 (1992)
7. Tange, H.J., Schouten, H.C., Kester, A.D.M., Hasman, A.: The granularity of medical narratives and its effect on the speed and completeness of information retrieval. Journal of the American Medical Informatics Association 5(6), 571 (1998)
8. Brown, P.J.B., Sönksen, P.: Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model. Journal of the American Medical Informatics Association 7(4), 392 (2000)
9. Sullivan, F., Gardner, M., Van Rijsbergen, K.: An information retrieval service to support clinical decision-making at the point of care. The British Journal of General Practice 49(449), 1003 (1999)
10. Noone, J., Warren, J., Brittain, M.: Information overload: opportunities and challenges for the gp's desktop. Studies in Health Technology and Informatics 52, 1287 (1998)
11. Gell, G., Oser, W., Schwarz, G.: Experiences with the aura free text system. Radiology 119, 105–109 (1976)
12. Gell, G.: Aura: routine documentation of medical texts. Methods Inf. Med. 22, 63–68 (1983)
13. Zingmond, D., Lenert, L.A.: Monitoring free-text data using medical language processing. Computers and Biomedical Research 26(5), 467–481 (1993)
14. Holzinger, A., Geierhofer, R., Errath, M.: Semantische informationsextraktion in medizinischen informationssystemen. Informatik-Spektrum 30(2), 69–78 (2007)
15. Gregory, J., Mattison, J.E., Linde, C.: Naming notes: transitions from free text to structured entry. Methods of Information in Medicine 34(1-2), 57 (1995)
16. Holzinger, A., Kainz, A., Gell, G., Brunold, M., Maurer, H.: Interactive computer assisted formulation of retrieval requests for a medical information system using an intelligent tutoring system. In: Proceedings of ED-MEDIA, pp. 431–436 (2000)
17. Lovis, C., Baud, R.H., Planche, P.: Power of expression in the electronic patient record: structured data or narrative text? International Journal of Medical Informatics 58, 101–110 (2000)
18. Smalheiser, N.R., Swanson, D.R.: Using arrowsmith: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57(3), 149–153 (1998)

19. Harter, S.P., Hert, C.A.: Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods. Annual Review of Information Science and Technology (ARIST) 32, 3–94 (1997)
20. Kreuzthaler, M., Bloice, M.D., Simonic, K.M., Holzinger, A.: On the Need for Open Source Ground Truths for Medical Information Retrieval Systems. In: International Conference on Knowledge Management and Knowledge Technologies, vol. 10, pp. 371–381 (September 2010)
21. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Setzer, A., Roberts, I.: Semantic Annotation of Clinical Text: The CLEF Corpus. In: Workshop Programme, p. 19 (2008)
22. Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D.J., Johnson, N., Cohen, K.B., Duch, W.: A shared task involving multi-label classification of clinical free text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, pp. 97–104. Association for Computational Linguistics (2007)
23. Hersh, W.R., Müller, H., Jensen, J.R., Yang, J., Gorman, P.N., Ruch, P.: Advancing biomedical image retrieval: development and analysis of a test collection. Journal of the American Medical Informatics Association 13(5), 488 (2006)
24. Müller, H., Deselaers, T., Deserno, T.M., Clough, P., Kim, E., Hersh, W.: Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks. In: Peters, C., Clough, P., Gey, F.C., Karlgren, J., Magnini, B., Oard, D.W., de Rijke, M., Stempfhuber, M. (eds.) CLEF 2006. LNCS, vol. 4730, pp. 595–608. Springer, Heidelberg (2007)
25. Ogren, P.V., Savova, G., Buntrock, J.D., Chute, C.G.: Building and Evaluating Annotated Corpora for Medical NLP Systems. In: AMIA Annual Symposium Proceedings, p. 1050. American Medical Informatics Association (2006)
26. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J.S., Roberts, I., Setzer, A., Tapuria, A., et al.: The CLEF corpus: semantic annotation of clinical text. In: AMIA Annual Symposium Proceedings, p. 625. American Medical Informatics Association (2007)
27. Uzuner, Ö., Luo, Y., Szolovits, P.: Evaluating the State-of-the-Art in Automatic De-identification. Journal of the American Medical Informatics Association 14(5), 550 (2007)
28. Uzuner, Ö., Goldstein, I., Luo, Y., Kohane, I.: Identifying patient smoking status from medical discharge records. Journal of the American Medical Informatics Association 15(1), 14–24 (2008)
29. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval. Addison-Wesley, Reading (1999)
30. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval. Cambridge Univ. Pr. (2008)
31. Saracevic, T.: Evaluation of evaluation in information retrieval. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 138–146. ACM (1995)
32. Wilson, T.D.: Human information behavior. Informing Science 3(2), 49–56 (2000)
33. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) 2(1), 1–19 (2006)

34. Kreuzthaler, M., Bloice, M.D., Faulstich, L., Simonic, K.-M., Holzinger, A.: A comparison of different retrieval strategies working on medical free texts. Journal of Universal Computer Science 17(7), 1109–1133 (2011)
35. Wingert, F.: Automated indexing based on SNOMED. Methods of Information in Medicine 24(1), 27–34 (1985)
36. Wingert, F.: Morphologic analysis of compound words. Methods of Information in Medicine 24(3), 155 (1985)
37. Wingert, F.: An indexing system for SNOMED. Methods of Information in Medicine 25(1), 22–30 (1986)