MedEval — A Swedish Medical Test Collection with Doctors and Patients User Groups

Karin Friberg Heppin
Department of Swedish
University of Gothenburg
Gothenburg, Sweden
karin.friberg@svenska.gu.se

Abstract

MedEval is a Swedish medical test collection where assessments have been made, not only for topical relevance, but also for target reader group: Doctors or Patients. The user of the test collection can choose if s/he wishes to search in the Doctors or the Patients scenarios where the topical relevance assessments have been adjusted with consideration to user group, or to search in a scenario which regards only topical relevance. MedEval makes it possible to compare the effectiveness of search terms when it comes to retrieving documents aimed at the different user groups. MedEval is also the first medical Swedish test collection.

1 A New Test Collection

When the decision was made to build a new test collection, the Department of Swedish at the University of Gothenburg was involved in projects of research in medical language processing. There was also a growing interest of research in information retrieval. There existed no Swedish medical test collection. Creating one seemed to be a good investment in knowledge and resources, even though this involved a team of people during many months. As building a test collection is a major undertaking not many exist. OHSUMED is a medical test collection, albeit in English. It is built on nearly 350 000 references from MEDLINE. The OHSUMED documents are assessed on a three graded scale: definitely, possibly and not relevant. OHSUMED contains 106 topics generated by physicians from authentic situations. The topics consist of both information about the patient and the request. (OHSUMED, 2007)

With a new collection such as MedEval, the Swedish department could take control over the architecture and make decisions such as using a four graded scale of relevance, making it possible to employ a variety of evaluation tools. However, the most important decision was to assess documents, not only for relevance to topics, but also for intended groups of readers, 'Doctors: medical professionals' or 'Patients: lay persons', and to allow the user to choose user scenario: None, Doctors or Patients.

2 Documents

The MedEval test collection is built on documents from the MedLex medical corpus (Kokkinakis, 2004). MedLex consists of scientific articles from medical journals, teaching material, guidelines, patient FAQs, health care information, etc. The set of documents used in MedEval is a snapshot of MedLex in October 2007, approximately 42 200 documents or 15 million tokens (see table 1). The documents are stored in the trectext format.

3 Indexes

The MedEval test collection has two indexes. One where the documents are converted to lower case, tokenized and lemmatized, and one where the compounds also are decomposed. In the second index, the compound terms are indexed as a whole together with the compound constituents. For instance: the compound *saltkoncentration* 'salt concentration' is indexed as *saltkoncentration*, *salt*, and *koncentration*.

Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents, pages 1–7,
Los Angeles, California, June 2010.

Type of source	Number of documents	Percent of documents	Number of tokens	Percent of tokens
Journals and periodicals	8 453	20.0	5.3 million	34.6
Specialized sites	14 631	34.6	2.9 million	19.1
Pharmaceutical companies	9 200	21.8	2.3 million	14.8
Government, faculties, institutes, and hospitals	2 955	7.0	2.0 million	13.3
Health-care communication companies	4 036	9.6	1.7 million	11.3
Media (TV, daily newspapers)	2 980	7.1	1.0 million	6.9
Total	42 255	100.1	15.2 million	100

Table 1: The genres of the MedEval document sources. The document collection is a snapshot of the MedLex corpus in October 2007. (D. Kokkinakis, p.c.)

4 Topics

Two medical students in their fourth year of studies were hired to create the topics. Their instructions were to create information needs that could be requested in real medical situations. 100 topics were created in the first stage. 62 of these were used in the collection.

A topic consists of a title, a description and a narrative. The title is a short phrase summarizing the information need. The description is concise information about the topic, usually in the form of a question or a request. The narrative is a few sentences long and it stipulates what makes a document relevant to the topic. The narrative contains the guidelines for the assessors when judging the relevance of the documents in the next stage. An example of a topic is given below. The English equivalent of the description of topic 51 is: *Why can a patient with cancer contract anemia?*

<TOP>
<TOPNO>51</TOPNO>
<TITLE> Anemi och cancer </TITLE>
<DESC> Varför kan en patient med cancer drabbas av anemi? </DESC>
<NARR> Relevanta dokument ska innehålla information om vad anemi /blodbrist är, symtom, behandling och orsaker. Information om cancerrelaterad anemi dels utlöst av cancern och dels utlöst av cancerbehandlingen är relevant. </NARR>
</TOP>

5 Selecting Documents to Assess

In the ideal test collection every document would be assessed for relevance with respect to every topic. But with over 42 000 documents and 62 topics, taking 8 minutes to assess each document, it would take four persons more than 40 years working 40 hours per week to finish the assessments.

Instead, only the documents that were considered most likely to be relevant to each topic were assessed. The documents were filtered out by use of four queries, one specific and one exhaustive for each index. The documents selected for each topic were sorted by document ID and duplicates were removed. This was done so that the assessors would not know how high a document had been ranked, or in how many searches it had been retrieved. For each topic and each of the four queries the 100 highest ranked documents were selected, if, in fact, there were that many.

6 Relevance Judgments

For the relevance judgments four new medical students were consulted. For each of 62 topics, an assessor read through the documents to be assessed and decided, for each document, the intended group of readers and the degree of relevance to the topic. The documents for each individual need were assessed by one and the same assessor for reasons of consistency.

The MedEval relevance assessments were made on a four graded scale, 0-3, where 0 is 'Not at all relevant' and 3 is 'Highly relevant'. The scale is easily turned into a binary scale by stating that the documents with the lower grades are to be consid-

ered non-relevant and the ones with higher grades relevant. Where the division is made between relevant and non-relevant depends on the needs of the user in each case.

The relevance considered by the assessors was topical relevance, how well a document corresponds to a topic. The assessors were instructed not to involve user relevance in this score. Each document was judged on its own merits. The novelty of the contents of a document should not be considered.

7 Target Groups

In addition to topical relevance the assessors judged each document for reader target group, that is which group of readers was the intended: Patients, if a document was written for lay persons, or Doctors, if it was written for medical professionals.

For a classification of documents according to intended reader group to be useful, there must be a measureable difference between the document classes. Table 2 shows a number of type/token frequencies in different subsets of the collection. In each set duplicates were removed in the case that a document had been assessed for more than one topic. The subsets considered are described below. Full form types are the original terms of the documents before lemmatization and lemma types are the same terms after lemmatization.

Entire collection All documents of the MedEval collection.

Assessed documents All documents that have been assessed for any topic.

Doctors assessed All documents that for at least one topic have been assessed to have target group Doctors.

Patients assessed All documents that for at least one topic have been assessed to have target group Patients.

Common files All documents that for at least one topic have been assessed to have target group Doctors and for another to have target group Patients.

Doctors relevant All documents that for at least one topic have been assessed to have at least relevance grade 1 and to have target group Doctors.

Patients relevant All documents that for at least one topic have been assessed to have at least relevance grade 1 and to have target group Patients.

Before counting frequencies, the files were cleaned from tags, IDs, dates (in the date tag, not in the actual text), web information and punctuation marks. Some observations are readily made by studying table 2.

The number of tokens per document is significantly smaller for the entire collection, than for any subset. This means that there is a large number of short documents that were not retrieved by any query when the documents to be assessed were selected. Maybe not surprising, since short documents contain few terms which can match the queries.

The documents in the set 'Patients assessed' had only 57% the number of tokens per document, compared to the documents in 'Doctors assessed'. Even though there were over 1 000 more documents in 'Patients assessed' than in 'Doctors assessed', there were over 50 000 more lemma types in the doctor documents and almost 30 000 more lemma compound types. The average word length in 'Doctors assessed' was 6.29 compared to 5.73 for 'Patients assessed'. The ratio of compound tokens was also higher in the doctor documents, 0.128 compared to 0.098.

Table 3 illustrates the fact that the doctor documents contain more and longer terms and more compounds than patient documents. This table shows frequencies of all full form types of strings beginning with *förmak* 'atria' in 'Patients assessed' and 'Doctors assessed' respectively. The patient documents have 18 full form types beginning with *förmak* while doctor documents have 75. That is more than four times more types for the doctor documents.

A closer look at the frequencies of *förmak*$*$ in the professional and lay person texts reveals that not all frequencies are higher for professionals. The frequencies of nouns in the definite form in the lay person texts are close to, equal or higher than the same forms in the professional texts.

	Entire collection	Assessed documents	Doctors assessed	Patients assessed	Common files	Doctors relevant	Patients relevant
Number of documents	42 250	7 044	3 272	4 334	562	1 233	1 654
Tokens	12 991 157	5 034 323	3 232 772	2 431 160	629 609	1 361 700	988 236
Tokens/document	307	715	988	561	1 120	1 104	596
Average word length	5.75	6.04	6.29	5.73	6.16	6.33	5.63
Full form types	334 559	181 354	154 901	92 803	50 961	87 814	43 825
Lemma types	267 892	146 631	126 217	73 121	40 857	71 974	34 263
Compound tokens	1 273 874	573 625	412 475	237 267	76 117	179 580	92 420
Full form compound types	187 904	99 614	83 846	47 387	24 083	45 257	20 157
Lemma compound types	144 159	78 508	66 907	37 151	19 685	36 867	16 006
Ratio of compounds	0.098	0.114	0.128	0.098	0.120	0.132	0.094

Table 2: Type and token frequencies of the terms in different subsets of the MedEval test collection.

Looking at all instances of strings beginning with *förmak∗* in the two sets of documents there is a significant difference. In the patient documents 66 tokens of 372, or 17.7%, are nouns in the definite form, while the corresponding numbers for the doctor documents is 89 of 932 tokens, or 9.6%. At this stage one can only speculate why this is so. A hypothesis is that doctors/medical professionals often discuss matters in a generic point of view, while patients/lay persons discuss specific cases.

Term	Doctors	Patients
förmaken	21	21
förmakens	1	2
förmaket	11	14
förmaksflimret	16	28
förmaksmyocyterna	2	1

Table 4: Frequencies of terms beginning with *förmak* 'atria', which are in the definite form in the set 'Patients assessed'. The frequencies of these word forms in the documents written for the two target groups are compared.

8 User Groups

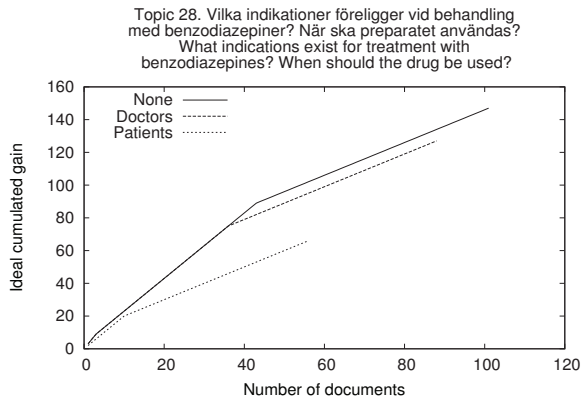
The MedEval test collection allows the user to state user group: *None* (no specified group), *Doctors* or *Patients*. This choice directs the user to one of three scenarios. The None scenario contains the topical relevance grades as made by the assessors. The Doctors scenario contains the same grades with the exception that the grades of the documents marked for Patients target group are downgraded by one. In the same way the Patients scenario has the documents marked for Doctors target group downgraded by one. This means that for a doctor user patient documents originally given relevance 3, are graded with 2, documents given relevance 2 are graded 1 and documents given relevance 1 are graded 0. The same is done in the Patients scenario with the doctor documents. The idea is that a document that is written for a reader from one target group but retrieved for a user from the other group will not be non-relevant, but less useful than a document from the correct target group. Put differently, a document intended for patients would contain information that doctors (hopefully) already know. On the other hand, documents intended for doctors, even though they might be topically relevant for a patient's need, run a great risk of being written in such a way that a patient will have problems grasping the whole content.

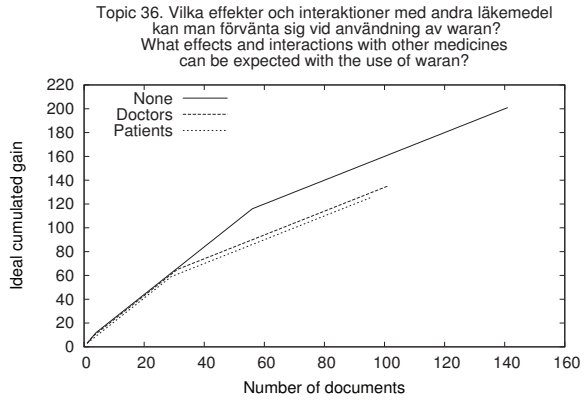
Adjusting relevance in the manner described affects the scenario recall bases. Since relevance grades are downgraded for documents of the opposing target group there will be fewer relevant documents in the Doctors and Patients scenarios than in the None scenario. This is demonstrated in figure 1 where the ideal cumulated gain for the three scenarios of topics 28, 36 and 92 are shown. The ideal cumulated gain is the maximum score of retrieved information possible at each position in a ranked list of documents (Järvelin, Kekäläinen, 2002). The score for each position is the sum of all relevance scores so far in the ranked list.

The three topics of figure 1 show different characteristics with reference to the number of relevant

Lay person audience	förmak	73	förmaksflimmer	219
	förmaken	21	förmaksflimmerattacker	1
	förmakens	2	förmaksflimmerpatienter	1
	förmaket	14	förmaksflimret	28
	förmaks	1	förmakslimmer	1
	förmaksarytmier	2	förmaksmyocyterna	1
	förmakseffekt	1	förmakstakykardi	1
	förmaksfladder	2	förmaksutlösta	2
	förmaksflimer	1	förmaksöra	1
Professional audience	förmak	93	förmaksmuskelns	1
	förmaken	21	förmaksmuskulaturen	2
	förmakens	1	förmaksmyocyterna	2
	förmaket	11	förmaksmyokard	3
	förmakets	1	förmaksmyokardiet	1
	förmaks	21	förmaksmyxom	2
	förmaksaktivering	1	förmaksnivå	2
	förmaksaktivitet	1	förmaksnära	1
	förmaksaktiviteten	2	förmaksoch	1
	förmaksanatomi	1	förmakspacing	7
	förmaksarytmi	2	förmakspeptider	1
	förmaksarytmier	9	förmaksrytmer	1
	förmaksbidraget	1	förmaksseptostomi	1
	förmaksbradyarytmi	1	förmaksseptum	2
	förmaksdefibrillator	2	förmaksseptumaneurysm	10
	förmakseffekt	2	förmaksseptumdefekt	5
	förmaksfladder	57	förmaksseptumdefekten	1
	förmaksfladdret	2	förmaksseptumdefekter	1
	förmaksflimmer	544	förmaksseptums	1
	förmaksflimmerablationer	2	förmaksstimulerat	1
	förmaksflimmerattacker	1	förmaksstimulerin	5
	förmaksflimmerduration	2	förmaksstorlek	2
	förmaksflimmerepisoder	4	förmaksstorleken	1
	förmaksflimmerfladder	2	förmakssynkron	1
	förmaksflimmerpatienter	4	förmakssystole	1
	förmaksflimmerrecidiv	1	förmakstaket	1
	förmaksflimmertendensen	1	förmakstakykardi	11
	förmaksflimmerunderhållande	1	förmakstakykardie	8
	förmaksflimret	16	förmakstromb	2
	förmaksflimrets	4	förmakstryck	1
	förmaksfrekvenser	1	förmakstrycket	1
	förmaksfunktion	1	förmaksvolym	2
	förmaksförstoring	1	förmaksvägg	1
	förmaksimpuls	1	förmaksväggarna	2
	förmaksinhiberad	1	förmaksväggen	6
	förmakskontraktion	4	förmaksvävnaden	2
	förmakskontraktionen	6	förmaksöra	9
	förmakskontraktionens	1	förmaksöronen	2
	förmaksmuskeln	1		

Table 3: This is a randomly chosen example of the difference in the number of types and of tokens in the documents written for a lay person audience, in the set 'Patients assessed' and the ones written for a professional audience, in the set 'Doctors assessed'. The table shows all types of strings beginning with *förmak* 'atria' in documents written for the two target groups. The number of tokens for each type is also shown.





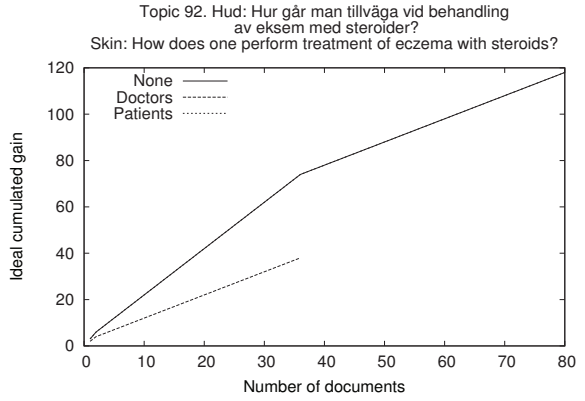


Figure 1: The recall bases of topic 28, 36 and 92 represented in ideal cumulated gain for the three scenarios: None, Doctors and Patients. For topic 28 most of the highly relevant and fairly relevant documents were assessed to have target group Doctors. Topic 36 had the relevant documents spread fairly evenly between the Doctors and Patients target groups. Topic 92 showed no documents of any relevance grade for documents marked for target group Doctors. Thus the None and the Patients ideal gain vector coincide fully, while the cumulated gain for the Doctors scenario is very low.

doctor and patient documents. Topic 36 has fairly similar cumulated gain curves for the Doctors and Patients scenarios. Topic 28 has a majority of doctor documents, while topic 92 had no documents of any relevance grade for documents marked for target group Doctors. Thus the None and the Patients ideal gain vector coincide fully, while the cumulated gain for the Doctors scenario is very low, originating from downgraded patient documents.

9 Example Runs

To demonstrate the effectiveness of search terms from the different styles of language of the two target groups, the synonyms *anemi* 'anemia' and *blodbrist* 'blood lack' were run as search keys for topic 51 in the Doctors and Patients scenarios. *anemi* is a neoclassical term, belonging to the professional language and *blodbrist* is the corresponding lay person term.

In the Doctors scenario the difference between the results of the two search keys was striking: full recall for the neoclassical term quite early in the ranked list of documents and no recall at all for the lay person term. The Patients scenario did not show as big difference between the search keys. Note that the resulting ranked lists of documents is the same for both scenarios for the same search key. It is the relevance grades of the retrieved documents that differ.

Scenario	Recall	anemi	blodbrist
Doctors	@10	50% (4/8)	0% (0/8)
	@20	100% (8/8)	0% (0/8)
	@100	100% (8/8)	0% (0/8)
Patients	@10	22% (4/18)	33% (6/18)
	@20	39% (7/18)	39% (7/18)
	@100	66% (12/18)	56% (10/18)

Table 5: Running the synonyms *anemi* 'anemia' and *blodbrist* 'blood lack' as search keys for topic 51 in the Doctors scenario gave full recall early in the ranking list for the neoclassical term *anemi*, but no recall at all for the lay person term *blodbrist*. In the Patients scenario the difference in effectiveness for these search keys was not as striking.

10 Final Words

This paper shows a few aspects of medical information retrieval which can be studied with the use of the MedEval test collection. The main novelty of the collection is the marking of document target groups, Doctors and Patients, together with with the possibility to choose user group. This opens up new areas of research in Swedish information retrieval such as how one can retrieve documents suited for different groups of users.

The Department of Swedish at the University of Gothenburg is in the process of making the MedEval test collection available to academic researchers.

Acknowledgments

The author would like to thank the FIRE (Finnish Information Retrieval Experts) research group at the University of Tampere, Finland, for their invaluable help in building the MedEval test collection.

References

OSHUMED. 2007. *The OHSUMED test collection*. [www] <http://ir.ohsu.edu/ohsumed/ohsumed.html>.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, Vol. 20, No.4, pages 422-446.

Dimitrios Kokkinakis. 2004. *Medlex: Technical report*. Department of Swedish, University of Gothenburg, Sweden. [www] <http://demo.spraakdata.gu.se/svedk/pbl/MEDLEX_work2004.pdf>.