MedEval
Six test collections in one

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

Abstract

Information retrieval is a field of research reaching from computer and information science to lingusitics. As a linguist in the information retrieval field, I leave the quest for effective search engines and evaluation models to others, and focus on language aspects. Words, and parts of words such as compound constituents, which are successful in queries, what features do they have in common? Does the domain of search terms have impact in a domain specific environment? Can search terms with certain features help users of different categories find documents suited for them?

This paper describes the making of an information retrieval test collection which made it possible to study these questions. The test collection will be used to **Eval**uate search strategies to retrieve **Med**ical documents, hence the name.

To study language aspects of information retrieval a new test collection was called for, a collection which was domain specific, which regarded user groups, and which had double indexes for split and unsplit compounds. Since there was no such collection we built **MedEval**, a Swedish medical test collection, with documents marked for target groups, professionals and laypersons, with a system allowing choice of user group, and with two indexes, treating compounds in different ways.

In accordance with the Cranfield Paradigm the MedEval test collection is based on three parts: A set of **documents**, a set of **topics**, and a set of known **relevant documents** with respect to each of the topics (Cleverdon, 1967).

1 The Document Collection

The MedEval test collection is built on documents from the MedLex 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 15.2 million tokens. See Table 1.

For the MedEval test collection the documents are stored in the trectext format. The documents have IDs that reveal the source, and they are tokenized and tagged.

Table 1: The genres of the MedEval document sources. (D. Kokkinakis, p.c.)

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
Faculties, institutes, hospitals and government	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

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 223–226

2 The Indexes

The terms of the documents and their positions in each document are listed in inverted files. For each term, the ID of each document containing this term is listed along with the positions of the term in the document. This makes it possible to search for phrases or put conditions on queries, for example that terms must appear in a certain order or within a certain distance of each other.

The MedEval test collection has two indexes. One that contains the documents converted to lower case, tokenized and lemmatized, and one that also has compounds split before lemmatization. The compounds are indexed as one orthographic word, as in the first index, but also by each part separately. For example *spiralformad* 'spiral formed', indexed as *spiralformad*, *spiral*, and *formad*. Example 1 shows part of a document prepared for the first index. Example 2 is the same text prepared for the second index, with split compounds.

Example 1. A document prepared for the first index. It is tagged and the words are converted to lower case, tokenized and lemmatized.

<DOC>
<DOCNO> FLKB-0004 </DOCNO>
<TITLE> cell vävnad kropp organisation </TITLE>
<DATE> 2006-04-xx </DATE>
<TEXT> http://www.folkbildning.net ...
senare uppstå dna deoxiribonukleinsyra en spiralformad molekyl uppbyggd av kolhydrat fosfat och kvävebas det vara också möjlig att de första dna-molekyl sprida som ett smittämne från någon annan plats i rymd där levande organism redan finna för att cell skola överleva och dessutom trivas vara det viktig att miljö ...

</TEXT>
</DOC>

Example 2. A document prepared for the second index. The text contains the compounds as a whole, as well as the parts.

<DOC>
<DOCNO> FLKB-0004 </DOCNO>
<TITLE> cell vävnad kropp organisation </TITLE>
<DATE> 2006-04-xx 2006-04- xx </DATE>
<TEXT> http://www.folkbildning.net ...
senare uppstå dna deoxiribonukleinsyra deoxiribo nuklein syra en spiralformad spiral formad molekyl uppbyggd upp byggd av kolhydrat kol hydrat fosfat och kvävebas det vara också möjlig att de första dna-molekyl dna- molekyl sprida som ett smittämne smitt ämne från någon annan plats i rymd där levande organism redan finna för att cell skola överleva och dessutom trivas vara det viktig att miljö ...

</TEXT>
</DOC>

3 Topics

When the documents were assessed, it was the relevance of a document in accordance to a topic that was judged. The topics are static and are used as a base for posing queries. Queries, on the other hand, are created by the user and put to the system in order to find documents that satisfy the topic. They are specific for each run and can be modified if the user is not satisfied with the results.

The process of developing topics, which is described below, is inspired by INEX 2006 Guidelines for Topic Development (Larsen and et al., 2006).

Two medical students were consulted to create topics. They were instructed to make the topics models of realism, sufficiently abstract to be assessed by others. The topics should have varying but suitable numbers of relevant documents, not lower than 5 and possibly up to 50 or more.

The topic creators made queries to the system to get an indication of the amount of relevant documents. Too many hits is an indication that the query is too general. With too many hits there is no room to test strategies for possible improvements. Of course it is equally important to check that relevant documents do exist.

The next step was to explore the collection again, more thoroughly, to see if the topics were suitable to enable assessors to consistently judge and grade documents for relevance. These trial runs helped the creators to decide the complexity of the topics.

When the main idea and title of a topic were ready, the narrative was constructed. The narrative explains in detail what makes a document relevant. It was the narratives that the assessors later used as guides when deciding the grade of relevance of the documents.

After the narrative, the description, in essence the topic itself, was written. A description is a natural language interpretation of the topic, written in one or two sentences. It is usually in the form of a question or a request.

The topics were converted to XML format, just as the documents. Each topic is surrounded by tags and also contains tags for topic number, title, description, and narrative.

Example 3. Example of a topic with ID number, title, description and a more informative narrative.

<TOP>
<TOPNO> 23 </TOPNO>
<TITLE> Risker vid användning av neuroleptika <TITLE>
<DESC> Vilka risker är förknippade med användandet av neuroleptika? </DESC>
<NARR> Relevanta dokument skall innehålla generell information gällande neuroleptika, deras indikationer, biverkningar och behandlingsalternativ. Information om de olika sjukdomstillstånd där neuroleptika används för behandling är relevant.
</NARR>
</TOP>

4 Relevance Assessments

With the topics created, documents were assessed for relevance with respect to each topic. Four new medical students were consulted, as the topic creators could not stay on. 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 assessor began by studying a topic so that (s)he became familiar with it. (S)he was also instructed to keep a written copy of the need at hand when reading the documents. The assessor read every document carefully, marking, in the margins, paragraphs contributing to the topic. After reading a document through, the assessor looked through the marked paragraphs and decided which degree of relevance the document should be assigned.

Each document was judged on its own merits. That is, seeing a piece of information for the umpteenth time should not tempt the assessor to judge it less relevant to the topic than it was the first time.

In the MedEval test collection the **relevance assessments** were made on a four graded scale, 0-3, according to the recommendation by Sormunen (2002). See Table 2. Four levels, instead of the usual two, allow for a subtler differentiation in the evaluation of search strategies, when it comes to retrieval of highly relevant documents compared to moderately relevant documents. The scale is easily turned into a binary scale if one regards documents graded 0 or 1, as well as unassessed documents, as non-relevant and documents graded 2 or 3 as relevant. The relevance judged here is the **topical relevance**, how well the document corresponds to the topic. The assessors were instructed not to involve **user relevance** in this grade, that is how relevant a document is to a certain user at a certain point of time.

When assessing the documents for **target group** the assessors decided for each document which group of readers was the intended. The assessors marked the documents with a **P**, for patients, if a document was written for laymen, or with an **L**, for *läkare* 'doctor', if it was written for medical professionals. The assessors were forced to mark either a **P** or an **L**. The assumption is that doctors and patients could both have a certain, although not equal, interest in most documents. A third category including both doctors and patients would open up for the risk of having the majority of the documents ending up there.

The marking of target group was done to make it possible to evaluate search strategies, not only considering relevance to the topic, but also considering if the retrieved documents were aimed at the correct user profile.

5 Selecting Documents to Assess

In the ideal test collection every document would be assessed for relevance with respect to every topic. With a collection of 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 a series of queries using different strategies. The documents for each topic were sorted by document ID and duplicates were removed so that the assessors would not know how high a document had been ranked, or in how many searches it was retrieved. For each topic and each of the four search methods the 100 highest ranked documents were selected, if, in fact, there were that many. This means that for every topic between 100 and 400 documents were assessed. The mean number of assessed documents for a topic was 224, and the

Table 2: The four graded scale of topical relevance, according to Sormunen (2002).

Value	Relevance	Description
0	Non-Relevant	The document does not contain information relevant to the topic.
1	Marginally relevant	The document does not contain other information relevant to the topic than what is in the description of the topic.
2	Fairly relevant	The document contains more information about the topic than the description, but it is not exhaustive. If it is a topic with several aspects, only some aspects are covered.
3	Highly relevant	The document discusses all themes of the topic. If it is a topic with several aspects, all or most of them are covered.

mean number of documents judged relevant for a topic was 20. Selecting documents in this manner made the work load reasonable, but one must remember that all relevant documents may not have been assessed. Given more funding, we will in a later stage assess additional sets of document selected with other search engines and other queries.

6 Six Collections in One

The MedEval test collection allows the user to state **user group**: *None* (No specified group), *Doctor* or *Patient*. This choice directs the user to one of three scenarios. The *None* scenario contains the original relevance grades. The *Doctor* scenario contains the same grades with the exception that the grades of the documents marked for patient target group are downgraded by one. In the same way the *Patient* scenario has the documents marked for doctor 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 patient scenario with the doctor documents. The idea is that a document that is written for a reader from one user 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. More precisely, a document intended for a patient target group would (hopefully) contain background facts that most doctors already know. On the other hand, documents intended for the doctor target group, even though they might be topically relevant for a patient's need, the risk is that they are written in such a way that the patient has difficulty grasping the content.

In addition to indicating user group, the user must choose which index to search in, with or without split compounds. This choice is present for all three user scenarios. This means that the same query in connection with the same topic will give six different results depending on which user scenario and which index are chosen.

7 Using MedEval

A Swedish medical test collection such as MedEval with double indexes containing split and unsplit compounds, as well as the marking of document target group combined with the possibility to choose user group, will open up new linguistic aspects of Swedish information retrieval. How does one best deal with compounds? How does one get search results suited for different groups of users? And are there certain aspects to consider when searching in a domain specific environment.

Once the copyright issues are settled, we plan to let the MedEval collection be available to whomever wishes to use it.

References

C.W. Cleverdon. 1967. The cranfield tests on index language devices. In *Aslib Proceedings*, volume 19, pages 173–192.

Dimitrios Kokkinakis. 2004. Medlex: Technical report. Technical report, Department of Swedish Language, University of Gothenburg.

Birger Larsen and Andrew Trotman et al., 2006. *INEX 2006 Guidelines for Topic development*. [www] <http://inex.is.informatic.uni-duisburg.de/2006 /inex06/pdf/TD06.pdf>.

Eero Sormunen. 2002. Liberal relevance criteria of trec - counting on negligible ocuments? In *Proceedings of the 25^{th} Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206