MedEval
The Construction of a Swedish Medical Test Collection

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

1 Introduction

Finding documents with useful information is becoming increasingly difficult as the amount of information around us grows. Most of the existing search tools are constructed for documents in the English language. To make search engines for the Swedish language research needs to be done in Swedish. And to do that, resources in the Swedish language are necessary. For information retrieval that means test collections.

At Språkdata at the Department of Swedish Language, University of Gothenburg, I wanted to study information retrieval using a test collection having three features that, to my knowledge, did not occur in any existing Swedish test collection. I wanted the collection to be domain specific, more precisely, from the medical domain, to regard user groups, medical professionals and patients, and to give a possibility to study compounds, with double indexes for split and unsplit compounds.

Since I did not know of any such test collection, a decision was made to build **MedEval**, a test collection of Swedish medical documents, with two indexes, each treating compounds in a different way, and also with documents marked for target groups and a system giving a possibility to choose user group. The test collection is intended to be used to **Eval**uate search strategies to retrieve **Med**ical documents, hence the name.

2 The Document Collection - MedLex

MedEval is a test collection built on documents from the MedLex collection (Kokkinakis, 2004). MedLex is a workbench for lexicographic work, which consists of two main parts: a medical lexicon and a medical corpus.

The MedLex corpus consists of scientific articles from medical journals, teaching material, guidelines, patient FAQs, health care information, etc. The documents used in MedEval are documents collected for MedLex up until October 2007. The size is approximately 42 000 documents or 15 million tokens.

3 Documents

A test collection consists of three parts: A set of **documents**, a set of **information needs**, also called topics, and a set of **relevant documents** for each of the information needs.

For the MedEval test collection the documents from MedLex are stored in the trectext format. The documents are given an ID that reveals the source, and they are tokenized and tagged. A date is included if the date of the publication is known. If the document is from the Internet, the web address is supplied in the first line of the text.

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 the term is listed, along with the position of the term in the document. This makes it possible to search for phrases or to put conditions on queries, for example that terms must appear in a certain order or within a certain distance from each other.

The MedEval test collection has **two indexes**. One that contains the documents converted to lower case, tokenized and normalized, and one that also has compounds split before normalization. The compounds are indexed as a whole, but, in the second index, also by each part separately.

4 Topics and Assessments

To come up with realistic information needs, or topics, I consulted two medical students, in their

fourth year of studies. They were instructed to create topics with varying but suitable numbers of relevant documents, somewhere not lower than 5, preferably over 10 and possibly up to 50 or more. Once the information needs were created, documents were assessed for relevance for each need.

The topic creators were not able to stay on, so four new medical students were consulted as assessors. For each of 62 topics an assessor read through the documents to be assessed and decided, for each document, the degree of relevance to the topic in question and the intended group of readers. The documents for each individual need were assessed by one and the same assessor for reasons of consistency.

In the MedEval test collection the **relevance assessments** were made on a four graded scale, 0-3. This is easily turned into a binary scale if one regards documents graded 0 or 1, as well as unassessed documents, as irrelevant and documents graded 2 or 3 as relevant. Assessments of the documents were also made for intended readers, or **target groups**, patients or medical professionals. This feature makes it possible to evaluate search strategies, not only considering relevance to the topic, but also considering if the retrieved documents are intended for the given user profile.

In the ideal test collection every document would be assessed for relevance to every information need. But with a collection of 42 000 documents and 62 information needs, taking, on average, 8 minutes to assess each document, it would take four persons approximately 40 years to finish the assessments. Instead, only the documents that were most likely to be relevant to each information need were assessed. These documents were collected in **pools** of documents where documents had been filtered out by use of a series of simple queries using different strategies. In the case of MedEval I sorted out the 100 highest ranked documents from four different searches for the pool. The documents in the pool were sorted by document ID and duplicates were removed so that the assessors would not know how high a document was ranked, or in how many different searches it was retrieved.

5 Six Collections in One

The finished 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 the 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 the 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 entirely irrelevant, but probably less useful than a document that is read by a reader from the intended target group.

In addition to indicating user group, the user must choose which index to search in, with or without split compounds. This choice is present in 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.

A Swedish medical test collection such as MedEval will fill a need in domain specific information retrieval. The double indexes, with 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 aspects of information retrieval. Once the copyright issues are settled, I plan to let the MedEval collection be available to whomever wishes to use it.

References

Kokkinakis, Dimitrios. 2004. *MEDLEX: Technical Report*. Department of Swedish Language, Språkdata, University of Gothenburg. [www] <http://demo.spraakdata.gu.se/svedk/pbl/MEDLEX_work2004.pdf>. The reference created October 29, 2008.