Data linguistica 22

Resolving Power of Search Keys
in MedEval a Swedish Medical Test Collection with User Groups: Doctors and Patients

av Karin Friberg Heppin

Akademisk avhandling för filosofie doktorsexamen
i språkvetenskaplig databehandling,
som enligt beslut av humanistiska fakultetsnämnden vid
Göteborgs universitet kommer att försvaras offentligt lördagen den
9 oktober 2010 kl. 10.15 i Lilla hörsalen, Humanisten.

GÖTEBORGS UNIVERSITET
HUMANISTISKA FAKULTETEN

Göteborg 2010

TITLE: Resolving Power of Search Keys in MedEval
a Swedish Medical Test Collection with User Groups: Doctors and Patients
LANGUAGE: English
AUTHOR: Karin Friberg Heppin

Abstract

This thesis describes the making of a Swedish medical test collection, unique in its kind in providing a possibility to choose user group: doctors or patients. The thesis also describes a series of pilot studies which demonstrate what kind of studies can be performed with such a collection. The pilot studies are focused on search key effectivity: What makes a search key good, and what makes a search key bad?

The need to bring linguistics and consideration of terminology into the information retrieval research field is demonstrated. Most information retrieval is about finding free text documents. Documents are built of terms, as are topics and search queries. It is important to understand the functions and features of these terms and not treat them like featureless objects. The thesis concludes that terms are not equal, but show very different behavior.

The thesis addresses the problem of compounds, which, if used as search keys, will not match corresponding simplex words in the documents, while simplex words as search keys will not match corresponding compounds in the documents. The thesis discusses how compounds can be split to obtain more matches, without lowering the quality of a search.

Another important aspect of the thesis is that it considers how different language registers, in this case those of doctors and patients, can be utilized to find documents written with one of the groups in mind. As the test collection contains a large set of documents marked for intended target group, doctors or patients, the language differences can be and are studied. The author comes up with suggestions of how to choose search keys if documents from one category or the other are desired.

Information retrieval is a multi-disciplinary research field. It involves computer science, information science, and natural language processing. There is a substantial amount of research behind the algorithms of modern search engines, but even with the best possible search algorithm the result of a search will not be successful without an effective query constructed with effective search keys.

KEYWORDS: computational linguistics, natural language processing, information retrieval, test collection, medical language processing, resolving power, search keys, compounds

DISTRIBUTION:
Department of Swedish Language
University of Gothenburg
Box 200
SE-405 30 Gothenburg
Sweden

Data linguistica 22
ISSN 0347-948X
ISBN 978-91-87850-41-7

PRINTED in Sweden by Intellecta Infolog Göteborg 2010