Decomposition of Compounds and the Effect on Search Key Effectiveness in Information Retrieval

Karin Friberg Heppin

Språkbanken, Department of Swedish Language
University of Gothenburg
karin.friberg@svenska.gu.se

Abstract

Research on compound decomposition in information retrieval has mostly been performed comparing decomposition of all compounds in index and query with that of no compounds; and either all constituents have been used or none. This study suggests that improvements may be achieved by selective decomposition and by selective use of compound constituents.

1. Background

As a compound has at least two content-bearing morphemes and compounding is very productive in Swedish, a great part of information in Swedish text is contained in compounds. In the MedEval collection, used in this study, the ratio of compounds is 10% (Friberg Heppin, 2010).

In information retrieval it is a challenge to get at the information hidden in compounds. If a term in a document occurs only as a compound constituent, there will be no match in the search process if only the corresponding independent word is used as search key. A similar situation arises when a compound is used as search key, but only one or both constituents occur in the documents.

One approach to finding information in compounds is to decompose the compounds, that is split the compounds into constituents. Thereafter these can be added to the index or used in queries. When queries are expanded with compound constituents, recall may increase, but often the precision is also lowered. This may be the case if the compound is non-compositional or if the parts have high frequency and/or low specificity.

2. Experiments

This experiment was performed by running queries with one search key each for topic 1 (see figure 1) in the MedEval test collection. The measure used is normalized discounted cumulated gain, nDCG (Järvelin and Kekäläinen, 2002). Cumulated gain is calculated for each ranked position *i* by summing all relevance scores from 1 to *i*. A discounting factor reduces the amount of the score added for each step in the ranked list. Normalized DCG relates the achieved DCG to the maximum DCG possible for each position in the ranked list, the document cut off values.

```
<TOP>
<TOPNO>1</TOPNO>
<TITLE>The effects of a low-fat diet on
LDL and HDL </TITLE>
<DESC>How does a diet low in fat and
energy affect the concentrations of HDL
and LDL? </DESC>
<NARR> Relevant documents contain
information about the lipoproteins HDL and
LDL/light lipoproteins and their function,
and describes how a change in diet affects
the concentration of these in the blood.
Descriptions of low-fat diets are relevant.
</NARR>
</TOP>
```

Figure 1: Topic 1 of the MedEval test collection, here translated from Swedish into English.

2.1 The MedEval test collection

The MedEval test collection was built on documents from the MedLex medical corpus (Kokkinakis, 2004). MedLex consists of scientific articles from medical journals, teaching material, patient FAQs, health care information, etc. It has approximately 42 200 documents or 13 million tokens. The documents were assessed on a four-graded (0-3) scale of relevance. Two separate indexes were constructed: one where the document terms were lemmatized and one where the terms were lemmatized and compounds decomposed and indexed as a whole, together with the individual constituents.

2.2 The search keys

The terms examined were: the compound *fettsnål* 'fat stingy' (low-fat), the constituents *fett* 'fat', *snål* 'stingy' and the simplex word *kost* 'diet'. *fettsnål* could be classified as a derivation. If so, the second part would be a derivational affix which turns the stem noun into an adjective and diminishes it (Dura, 1998).

An essential difference between the terms *fett* 'fat' and *kost* 'diet', as used in topic 1, is that *fett* occurs as a constituent of another word while *kost* is an independent simplex word. The topic refers to a certain kind of diet, low-fat diet, however denoted by an adjective phrase. Before decomposition, the search key *fett* will not match instances of *fettsnål* from the topic, but *kost* as search key will match instances of *kost*.

What happens after decomposition is that we now match the search key *fett*, not only to *fett* as before, but also to the instances of the decomposed *fettsnål*, which was the term in the topic. We also match the term to other compounds which refer to the degree of fat, such as *fettintag* 'fat intake', *fettrik* 'fat rich' (rich in fat), and *fettinnehåll* 'fat content'. The search key *kost* will in turn have additional matches in compounds which have *kost* as a constituent. These compounds are mostly on a different level of specificity than the term *kost* from the topic. They are on the same level as *fettsnål kost* 'low-fat diet', but have a different meaning.

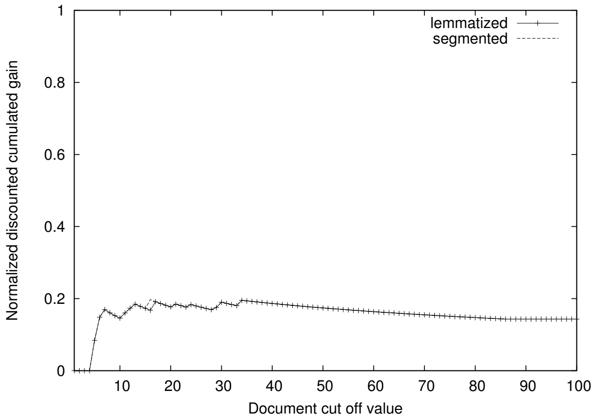

Figure 2: **fettsnål** 'fat stingy' (low-fat) – Somewhat effective. The term is essential for the topic but occurs in documents both relevant and not relevant to the topic. The nDCG curves for the non-decomposed and the decomposed index are almost identical. *fettsnål* is not itself common as a compound constituent, therefore it gives only a minimal difference between the indexes.

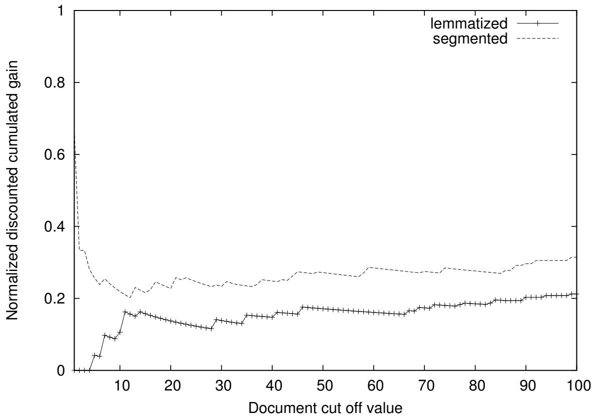

Figure 3: **fett** 'fat' – Moderately effective in the non-decomposed index but quite effective in the decomposed index. For both indexes the curve keeps rising slowly. This is a term which is general, has high frequency, and occurs in documents both relevant and not relevant to the topic.

3. Conclusions

The experiments show that compound decomposition in some cases improves the result, in some cases makes no difference and in other cases makes the result worse. This suggests that compounds and their constituents should not be treated equally. Instead selective decomposition should be employed when beneficial for the result.

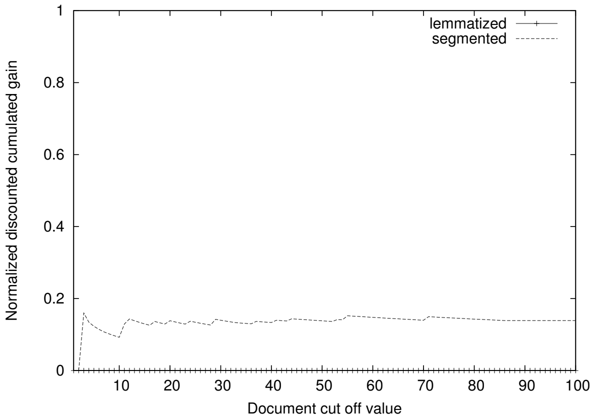

Figure 4: **snål** 'stingy' – This constituent is not relevant to the topic when used independently. There are no hits in the non-decomposed index, only in the decomposed index. This nDCG curve is similar, but slightly lower, to when the original term *fettsnål* was used. The documents retrieved with the search key *snål* are the same as the ones retrieved with the search key *fettsnål* (possibly a few more, see rank 2 and 3) but now mixed with additional noise, making the curve flatter.

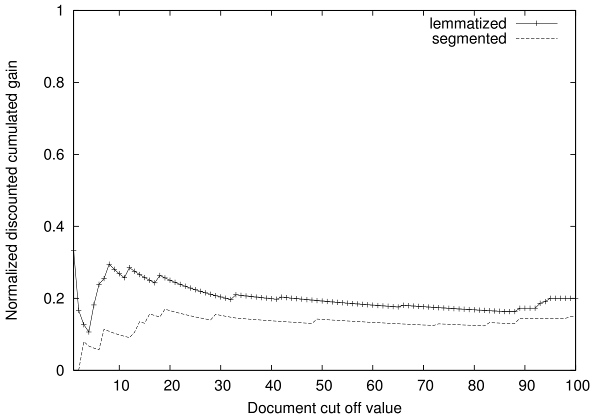

Figure 5: **kost** 'diet' – Somewhat effective. In the non-decomposed index the curve is throughout on a higher level than for the decomposed index. The search term is general, has high frequency and occurs in many documents, both relevant and non-relevant.

4. References

Elżbieta Dura. 1998. *Parsing words*. Ph.D. thesis, University of Gothenburg.

Karin Friberg Heppin. 2010. *Resolving Power of Search Keys in MedEval a Swedish Medical Test Collection with User Groups: Doctors and Patients*. Ph.D. thesis, University of Gothenburg. ⟨http://hdl.handle.net/2077/22709⟩.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Dimitrios Kokkinakis. 2004. MEDLEX: Technical report. Technical report, Department of Swedish, University of Gothenburg, ⟨http://demo.spraakdata.gu.se/svedk/pbl/MEDLEX_work2004.pdf⟩.