Query expansion using domain information in compounds

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

Abstract

This paper describes a query expansion strategy for domain specific information retrieval. Components of compounds are used selectively. Only parts belonging to the same domain as the compound itself will be used in expanded queries.

1 Introduction

Compounds are semantic units containing at least two content-bearing morphemes. They function as one word, and are, in many languages, written as one word. In Swedish newspapers around 10% of the words have been found to be compounds (Hedlund, 2002). Since a compound has at least two content-bearing morphemes, a great part of the information is contained in the compounds, information which can be essential in retrieving relevant documents.

I will study medical compounds, examining possible ways to expand queries in information retrieval using domain information. This information will guide the decision of when to include compounds parts in search queries. The theory is that components from the same domain as the compound itself, in this case the medical domain, will increase the effectiveness of the search, while components from other domains or standard language will not.

2 Information Retrieval

Information retrieval is about storing and organizing documents so that they can be found and retrieved when relevant to an information need (Baeza-Yates, and Ribiero-Neto, 1999). The words of the documents are stored in indexes. The user poses a query to the system containing words describing the information need. Words in the queries are matched against the indexed words. A ranking function finally ranks the documents in order of calculated relevance. The better the match, the higher the document is ranked.

The goal of information retrieval is to retrieve as many documents relevant to an information need as possible, **high recall**, and to have as low proportion of irrelevant documents in the output as possible, **high precision**.

2.1 Query expansion

Query expansion is modification of a query to improve retrieval effectiveness. This can be done by changing or increasing the term content of a query.

In my work the strategy of expanding queries, containing compounds, with selected compound components is discussed. The strategy should result in higher recall, since more documents are likely to be retrieved. There is, however, a risk of lower precision, since irrelevant documents with certainty also will be retrieved. To minimize the decrease of precision only components from the same domain as the compound itself will be used. Here, dealing with medical compounds, the objective is to decide if the components are from the medical domain.

3 Compounds

A compound is, as mentioned above, a semantic unit with more than one content-bearing morpheme. In Swedish, compounds are very productive. There is

an infinite number of possible compounds, so it is impossible to list them all. They are also written as one word without the boundary between the parts marked in any way.

3.1 Compositional/non-compositional compounds

Occasional compounds, constructed when needed, usually have a transparent meaning, where the meaning can be derived from the meaning of the parts. These are called **compositional compounds**. Other compounds, with a meaning that has strayed from the combined meaning of the components, are called **non-compositional compounds** (Hedlund, 2002). Non-compositional compounds are often lexicalized with a fixed meaning. An example of a lexicalized non-compositional compound is *trädgård* 'tree yard', Swedish for 'garden', not necessarily a garden containing trees.

In information retrieval, compositional and non-compositional compounds are best treated in different ways. Non-compositional compounds are often found in dictionaries and can be processed as they are. Using the components in queries would not benefit the result. If a compositional compound is used, the components might very well be used to expand the query, since they build up the meaning of the whole.

3.2 Decomposition not always beneficial

When expanding queries with compound components, to increase recall, it is important to be aware that this could result in lower precision. This might be the case if the compound is non-compositional or if the parts are too general or used in other domains. In Ahlgren (2004) the author gives examples of when decomposition of compounds is useful and when it is not. For a compound such as *fotboll* 'foot ball' (soccer), expanding a query with *fot* and *boll* would probably result in lower precision. On the other hand, expanding a query containing the compound *narkotikapolitik* 'drug politics', with *narkotika* and *politik* would probably be more useful. Documents containing phrases like *politik mot narkotika* 'politics against drugs' could be retrieved. Documents containing *narkotika* or *politik* alone would also be found. Here one can speculate that documents containing *narkotika* have a good chance of being relevant, while the concept *politik* is broader and could cause retrieval of many irrelevant documents.

My idea is to expand queries containing medical compounds by selecting components that also belong to the medical domain. Take the compound *korsband* 'cross band/tape' (cruciate ligament). Both parts belong to standard language. Including them would do more harm than good. In the case of *åderbråcksstrumpa* 'varicose-veins stocking' the component *åderbråck* seems to be a good candidate for query expansion, unlike *strumpa*, which belongs to standard language.

4 The Swedish MeSH thesaurus

The Swedish MeSH[1] (Svensk MeSH, www) is a medical thesaurus, a controlled vocabulary with words organized according to conceptual relations. A thesaurus can be used as a resource for indexing and searching in information retrieval. Choosing keywords in index and query according to such a controlled standard can lead to more documents being retrieved since the same term for a concept is used in both storage and search.

4.1 The MeSH tagger

A Swedish MeSH tagger (Kokkinakis, 2006) is being developed at Språkdata, Department of Swedish Language, Göteborg University. The tagger tags strings from six subdomains of the Swedish MeSH: **A**: Anatomy, **B**: Organisms, **C**: Diseases, **D**: Chemicals and Drugs, **E**: Analytical, Diagnostic, and Therapeutic Techniques and Equipment, and **F**: Psychiatry and Psychology. The tagger tags the longest found string from each subdomain. If a string is tagged, the tagger will not mark a substring of this from the same subdomain. The tagger does not tag any substrings shorter than five letters.

In the Swedish MeSH the compound *kransartär* 'wreath artery' (coronary artery) is not listed, thus it is not tagged. On the other hand *artär* is found and tagged accordingly. The word *krans* is not a medical term. It is not included in MeSH and consequently not tagged:

```
krans<mesh:A07>artär</mesh>
```

[1] The Swedish MeSH is based on a translation of the original American MeSH (Medical Subject Headings) (MeSH, www).

4.2 Expansion using MeSH

As mentioned above, one expansion strategy for queries with medical compounds is to add domain specific parts of the compounds to the query. This should work with compositional compounds. An example is *patellaluxation* 'patella dislocation' (dislocation of the knee cap). Chances are that documents containing any or both of the simplex words *patella* and *luxation* will be relevant to the needs of a user including *patellaluxation* in a query.

Baseline query:
`#sum(...patellaluxation...)`

Expanded query:
```
#sum (...#syn(patellaluxation
patella luxation)...)
```

Expanding queries with components not from the domain, especially those common in standard language, will probably result in lower precision. In the example *kransartär* the strategy would be to keep the original compound, add *artär* which is found by the MeSH tagger, but not *krans* which is not tagged.

Baseline query:
`#sum(...kransartär...)`

Expanded query:
`#sum(...#syn(kransartär artär)...)`

5 Experiments

To test the MeSH tagger, a run was made with 5 205 compounds extracted from the on-line medical lexicon Medlex (Kokkinakis, 2004), created at Språkdata, Department of Swedish Language, Göteborg University. Medlex was created by adding medical vocabulary to a learner's dictionary, thus a great part of the compounds in Medlex are from the medical domain.

895 of the 5 205 compounds were tagged. Among compounds not tagged, around 10% were medical. This figure should improve with a more comprehensive tagger.

233 compounds which were not tagged as a whole, had one or two components correctly tagged. This is where the strategy described should be most beneficial, suggesting that an expanded query contain the compound itself and the tagged substring(s).

Examples of tagging which may improve effectiveness in query expansion, are shown below:

`<mesh:D22/D27>cellgift</mesh>sbehandling`
'cell-poison treatment' (chemotherapy treatment)
`dotter<mesh:C04>tumör</mesh>`
'daughter tumor'
`fot<mesh:C17/C04/C02>vårta</mesh>`
'foot wart'
`<mesh:D06/D12>insulin</mesh>chock`
'insulin chock'

63 compounds had tagged components not used in medical senses. Those strings were homonymic, polysemic or had several facets. **Homonymy** is when a string represents different words that by chance are alike. **Polysemy** is when one word has several meanings. For example, the 'leg' of a person and the 'leg' of a table. **Facets** are different aspects of one concept. A 'person' has a body aspect as well as a personality aspect (Croft and Cruse, 2004).

It is tagging of words that are homonymic, polysemic, or with medical and non-medical facets that I predict will cause difficulties. An example is *hästansikte* 'horse face'. Although *ansikte* is a medical term, it is not used in a medical sense here. If you say that a person has a *hästansikte* it is a comment about looks, not health. The word *ansikte* has a medical facet, but also a personal appearance facet.

Other examples of compounds with problematic components are listed below:

`död<mesh:A02>skalle</mesh>`
'death skull' (skull referred to in a pirate or scary sense)
`femdygns<mesh:E01>prognos</mesh>`
'five-days prognosis' (weather domain)
`<mesh:A01>finger</mesh>borg`
'finger castle' (thimble)

Only four compounds were split incorrectly. An example is *röntgen+apparat* 'x-ray device', which was tagged as below, *nappar* meaning 'pacifiers':

`röntge<mesh:E07>nappar</mesh>at`

5.1 A pre-decomposed run

The MeSH tagger tags only the longest matching substring of a word in each of the six subdomains

of MeSH. If a compound is tagged as a whole, a component belonging to the same subdomain will not be tagged. The tagger also does not tag short strings unless they are separate words. This entails that short components will not be tagged unless decomposition of the compound is done first.

In order to see how these features affect the outcome of the tagger, I ran the Medlex list through the tagger after decomposing the compounds.

This time, 1 095 compounds had one or both components tagged. 819 of these were used in the medical sense. This is a number which should be compared with 233 in the previous run.

276 compounds had components that were tagged although the components were not used in a medical sense. Only one compound was split incorrectly.

5.2 Standard language versus medical language

One problem in decomposing compounds and using the medically tagged components to expand queries, is that many words that are medical in some meaning or facet are common in a standard language meaning or facet. Even if we know that such a component is used in the medical sense in a query, expanding the query with that component would bring on irrelevant documents. Examples of strings with such properties are *hand* 'hand' and *hjärta* 'heart'. Even though these words are used in medical senses they are also common in standard language, for example in lexicalized compounds or in phrases.

In the tagger run with the decomposed list, most of the 276 words that were tagged as medical, though not used in the medical sense, had as a component one of only 16 basic words. Below are a few such compounds that has as a component a word from that list, *hand*:

<mesh:A01>hand</mesh> bok
'hand book'
<mesh:A01>hand</mesh> broms
'hand brake'
<mesh:A01>hand</mesh> duk
'hand cloth' (towel)

6 Future work

I have presented a strategy of how to use domain information to decide when parts of a compound should be used in query expansion. Unfortunately I have not been able to test the effectiveness of this strategy. To get a true evaluation of the strategy, a Swedish medical test collection is needed. At present there is no such collection.

The first step is thus to create a Swedish medical test collection, the second to test query expansion strategies based on domain information, such as the one described here. The strategy described could be carried through not only in queries, but also in indexes. That is, if a document contains a medical compound, the index could contain not only the compound, but also its medical components. Still, a big challenge will be to work out how to deal with polysemic and homonymic words and words with medical and non-medical facets.

References

Ahlgren, Per. 2004. *The effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database*. Publications from Valfrid, nr 28. University College of Borås/Göteborg University.

Baeza-Yates Ricardo and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. ACM-press, New York, NY.

Croft, William and D. Alan Cruse. 2004. *Cognitive Lingustics*. Cambridge University Press. Cambridge.

Hedlund, Turid. 2002. Compounds in dictionary-based cross-language information retrieval. *Information Research*, volume 7 No.2. 2002. Department of Information Studies. University of Tampere, Finland.

Kokkinakis, Dimitrios. 2004. *MEDLEX: Technical Report*. Department of Swedish Language, Språkdata, Göteborg University. [www]. <http://demo.spraakdata.gu.se/svedk/pbl/MEDLEX_work2004.pdf> Retrieved January 9, 2007.

Kokkinakis, Dimitrios. 2006. Developing Resources for Swedish Bio-Medical Text Mining. *Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine (SMBM)*. Jena, Germany.

MeSH. *Medical Subject Headings*. U.S. National Library of Medicine, Bethesda, MD. [www]. <http://www.nlm.nih.gov/mesh/>. Retrieved January 9, 2007.

Svensk MeSH. *MeSH-resurser vid KIB*. Karolinska Institutet Universitetsbiblioteket, Stockholm. [www]. <http://mesh.kib.ki.se/>. Retrieved January 9, 2007.