Improving search engine retrieval using a compound splitter for Swedish

Hercules Dalianis
Department of Computer and System Sciences
KTH and Stockholm University
Forum 100, 164 40 Kista, Sweden
hercules@kth.se

Abstract

In this paper we have investigated 128 high frequent Swedish compound queries (6.2 per thousand) with no search results among 1.6 million searches carried out at nine public web sites containing all together 100,000 web pages in Swedish. To these compound queries we added a compound splitter as a pre-processor and we found that after decompounding these queries they gave relevant results in 64 percent of the cases instead of zero percent hits. We give also examples on some rules for optimal compound splitting in a search situation.

1 Introduction

Today when searching on Internet it is very likely that you will find some answer, this is due to the immense amount of information that is present and the efficient global search engines. There is always some web pages that contains the answer of your question, but when searching on a web site the task is not so easy anymore, the reasons are manifold.

One obvious reason is that one website contains much fewer web pages than the whole Internet, but other not obvious reasons are that the search engine on the website is lousy, this means it is slow or does not work, the index does not cover the whole website, and the hits are not relevant. The user search and does not get any hits, but the information must be there! What is wrong? We will here concentrate on the processing of the query that the user has entered into the search engine.

2 Previous research

The first thing that can happen in a query situation is that the user enters a search query to the search engine and he does not get any hit. This can be one of the 10-12 percent of the queries that is misspelled and hence does not give any matching to the search engine index. (Dalianis 2002, Sarr 2003). This can be solved using a spelling support linked to the index of the search engine. When the user makes a spelling error the spelling correction module tries to match a word that has either similar spelling or pronunciation to one or more words in the index and consequently the user will get feedback in form of possible candidate word(s), (Dalianis 2002, Google 2002, Sarr 2003, Stolpe 2003).

Of course the search word can be correctly spelled and the error can be in the web site but never the less we want to help the user to find the answer and we will also propose misspelled words. According to Dalianis (2002) around 90 percent of the spelling errors are corrected using a spelling correction algorithm.

Other problems in searching is often that the user searches for a word and the word is written in an other inflected form, this is of

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 38–42
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/
© by the author

course very common in cases when one uses languages that are morphologically complicated, (usually not English).

To solve the problem with word inflections one can use stemmers that will remove the inflections and make the words both in the search query and in the index stemmed and consequently able to match.

For Swedish, for example, precision and recall increased with 15 and 18 percent respectively using a stemmer, (Carlberger et al 2001). In Carlberger et al (2001) there is also an overview of different stemming experiments for European languages that show increase in precision and recall from 2-3 percent for English and up to 40 percent for Slovene, one can also read in Tomlinson (2003) that precision and recall increased immensely for European languages using stemming.

Two other methods to process queries are either compound splitting (decompounding) or compound joining. In Swedish for example we have a lot of compounds but we are heavily influenced by English written language and we tend to decompound Swedish words. This happens both in the situation when asking queries in search engines but also when writing text, therefore it is valuable to have a query analysis module that when obtaining sparse answers in a query situation tries to decompose the query word and consequently make a match possible. An other situation is when there are more than one search word and the user obtains no or sparse hits. Then the system should try different combinations in joining the words to compounds to obtain possible hits.

An example on the Swedish public medical website Vårdguiden, is when somebody is searching for *diabetespatient* and obtains no hits then the system tries to split the compound word to *diabetes patient* and the resulting hit become *patienter med diabetes* (patient with diabetes), notice that the stemmer will make it possible to automatically find the word *patienter* (plural form of patients). The other situation is that the user uses two search words *streptokock infektion* and does not obtain any hit then the system can propose the compound *streptokockinfektioner (plural form)* that gives several relevant hits.

A compound splitter/decompounder was used in Tomlinson (2003) and this gave good results in increasing precision and recall for Finnish and German but decreased precision and recall for other languages, Spanish, Dutch, French, Italian, Swedish and English.

Stemming and compound splitting was used in Chen & Gey (2004) they obtained 14 percent higher precision for Dutch, 37 percent for German and 30 percent higher precision for Swedish and Finnish respectively. Rosell (2003) obtained 10 percent better clustering results using compound splitting for Swedish when clustering Swedish texts

3 Our study and method

We have studied nine Swedish public websites they encompass two municipalities, one university, one political party, a nature conservation site, a public authority site, a popular science site, and two insurance companies, these web sites ranges the size from 500 documents to 50 000 documents. They contain totally 100 000 documents and the search engines there obtained around 1.6 million queries of which 9.3 percent were misspelled.

The top 30 of the total 1.6 million queries with no answer at all, were 6 000 compounds, 128 different compounds. In total 3.7 per thousand of the number of total queries. On some specific web sites there were up to 2 percent of the total queries has no answers. (Another 600 were written decompounded and became compounds by putting them together).

We can also estimate from the findings of Dalianis (2002) that 40 percent of the misspelled words were compound related, this should give up to 4 percent of the total amount of the problematic queries are compound related. This gives something up to 60 000 queries of the total 1.6 million queries would benefit of a compound splitter. Karlgren (2005) writes that around 10 percent of all words in Swedish running text are compounds.

We saw also that the two insurance company websites had a larger amount of compound queries in form of *studentförsäkring, skolförsäkring, garageförsäkring, villalarm, huslarm, hemlarm, bergvärme, luftvärmepump* (compounds with -insurance, -alarm, -heatpump) that does not give any hits without decomposition.

We connected the compound splitter described in (Sjöbergh & Kann 2004) to the search engine.

Proper nouns	**Ideal split (not carried out)**
Östrasjukhuset	Östra sjukhuset
Gothiacup	Gothia cup
Gröntkort	Grönt kort
Idrottenshus	Idrottens hus
Välacentrum	Väla centrum

Nouns	**Ideal split (not carried out)**
fossilabränslen	fossila bränslen
fenomenografi	fenomeno grafi
läs-och skriv svårigheter	läs- och skriv svårigheter

Table 2 and 3. The table shows five proper nouns and three nouns where the compound splitter failed.

Compound	**Oversplitting**	**Ideal split (not carried out)**
Mullvad	mull vad	mullvad
Helsingborgsdagblad	Helsingborgs dag blad	Helsingborgs dagblad
bilbarnstol	bil barn stol	bil barnstol
uppsatsdatabas	Uppsats data bas	uppsats databas
missbruksbehandling	miss bruks behandling	missbruks behandling
missbruksvård	miss bruks vård	missbruks vård
arbetskraftinvandring	arbets kraft invandring	arbetskraft invandring
arbetskraftsinvandrare	arbets kraft s invandrare	arbetskraft s invandrare
ordningsvaktsutbildning	ordnings vakts utbildning	ordningsvakts utbildning
gruppliv	grupp liv	gruppliv
Nattliv	natt liv	nattliv
Visakort	Visa kort	Visakort
luftvärmepump	luft värme pump	luft värmepump

Table 4. The table shows 13 compounds that became over split. All of the over split compounds have two parts shorter than 4 and 5 characters long respectively

We carried out compound splitting on each compound of the 128 compounds on each web site and it generated in total 7 724 new hits. 64 percent of them relevant to the query, 20 compounds were not splitted, over splitted or incorrectly splitted. That is 84 percent success rate of the compound splitter.

Of the 128 (100%) investigated compounds that none of the them obtained any hit at all first obtained hits after splitting them with a compound splitter and using the search engine on the split result again we found the following:

80 (64%) relevant hits boosting the search using compound splitting
29 (23%) gave us bad non relevant hits
+ 17 (13%) gave us still no answers
∑ 128 (100%)

One method to obtain good hits when searching is to only allow a certain distance

between the found decompounded parts of the original word. We would like to have some relations between the found words in the text. This can be carried out using the pseudo Boolean operator NEAR with say the parameter of 20 words distance.

How close to each other in the text should the splitted compound be to obtain relevant hits? We used the NEAR operator for counting number of words between the hits. The NEAR value could range from 1 to 70 words distance. Usually it was either around 1, 20 or 70 words distance. Average value 29 words distance.

4 Conclusions

Compound splitting as a post processing in a search engine works fine for Swedish, but one need a high quality compound splitter such that one does not get erroneous compound splitting that will deteriorate the precision of the search results. In other words bad compound splitting or over splitting will give us bad search results.

We have in our experiment seen that we obtained 64 percent more and relevant hits using the compound splitter described in (Sjöbergh & Kann 2004).

After our experiment we have also found that proper nouns need to be split in a smart and correct way. Nouns need to be split but not over split. We found also that a maximum of 29 words distance between the words in text in a compound splitting search gave relevant results. One clever strategy would then be that search hits using compound splitting should not stretch over sentence boundaries. 29 words can be considerd to be within one sentence distance.

Proper nouns need to be split in a smart and correct way. Nouns need to be split but not over split. One smart strategy is to split compounds at most two parts.

Hjelm and Schwarz (2005) that has been working with German compound splitting propose that The rightmost part of the compound part should be the longest. In Swedish we often use a genitive s in compounds. These “s” can be removed and the split boundary can be put there.

Some conclusions in rule form:

- Split compound in two parts
- Use genitive “s” as compound split marker and remove the “s”
- The rightmost should be the longest
- The compound split retrieval should be within one sentence, e.g. 29 words window.
- Treat Proper nouns specially.

Acknowledgements

I would like to thank Jonas Sjöbergh and Viggo Kann at KOD KTH for letting me use their compound splitter. I would also like to thank Johan Carlberger at Euroling AB for letting me use the statistics from the SiteSeeker search engine.

References

Carlberger, J., H. Dalianis, M. Hassel, O. Knutsson 2001. *Improving Precision in Information Retrieval for Swedish using Stemming*. In the Proceedings of NODALIDA 01 - 13th Nordic Conference on Computational Linguistics, May 21-22, Uppsala, Sweden.

Chen, A. and F. Gey. 2003. *Combining Query Translation and Document Translation in Cross Language Retrieval CLEF 2003*

http://clef.iei.pi.cnr.it/2003/WN_web/05.pdf

Cucerzan, S. and Eric Brill. 2004. *Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users*. In the Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, pp. 293-300. http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Cucerzan.pdf

Dalianis, H. 2002. *Evaluating a Spelling Support in a Search Engine*, in Natural Language Processing and Information Systems, 6th International Conference on Applications of Natural Language to Information Systems, NLDB 2002 (Eds.) B. Andersson, M. Bergholtz, P. Johannesson, Stockholm, Sweden, June 27-28, 2002. Lecture Notes in Computer Science. Vol. 2553. pp. 183-190. Springer Verlag.

Google. 2002. Search Engine Showdown: Google press release http://www.searchengineshowdown.com/newsarchive/000611.shtml (Visited June 7, 2005).

Hjelm, H. and C. Schwarz: *LiSa - morphological analysis for information retrieval*. In the Proceedings of Nodalida 2005 - 15th Nordic Conference on Computational Linguistics, May 21-22, Joensuu, Finland

Karlgren, J 2005. *Occurrence of compound terms andtheir constituent elements in Swedish*, In the proceeding of Nodalida 2005 - 15th Nordic Conference on Computational Linguistics, May 21-22, Joensuu, Finland

Rosell, M, 2003. *Improving Clustering of Swedish Newspaper Articles using Stemming and Compound Splitting*. In the proceeding of Nodalida 2003, the 14th Nordic Conference of Computational Linguistics, Reykjavik, May 30-31, 2003.

Sarr, M. 2003. *Improving precision and recall using a spell checker in a search engine*. In the proceeding of Nodalida 2003, the 14th Nordic Conference of Computational Linguistics, Reykjavik, May 30-31, 2003.

Sjöbergh, J. and V. Kann 2004. *Finding the correct interpretation of Swedish compounds, a statistical approach*, Proc. LREC 2004 (4th Int. Conf. Language Resources and Evaluation), Lissabon, Portugal. http://www.nada.kth.se/theory/projects/xcheck/rapporter/sjoberghkann04.pdf

Stolpe, D. 2003. *Högre kvalitet med automatisk textbehandling? En utvärdering av SUNETs Webbkatalog*, Examensarbete i Datalogi på Kungliga Tekniska Högskolan 2003. (Master thesis in Swedish)

Tomlinson, S. 2003. *Experiments in 8 European Languages with Hummingbird SearchServer™ at CLEF 2002*. To appear in Carol Peters, Martin Braschler, Julio Gonzalo and Michael Kluck, editors, Evaluation of Cross-Language Information Retrieval Systems: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Rome, Italy, September 19-20, 2002. Revised Papers. To be published by Springer in their Lecture Notes for Computer Science (LNCS) series.

http://www.stephent.com/ir/papers/clef02.html