

Evaluating a spelling support in a search engine

Hercules Dalianis

NADA-KTH

Royal Institute of Technology

100 44 Stockholm, Sweden

email: hercules@nada.kth.se

Abstract. The information in a database is usually accessed using SQL or some other query language, but if one uses a free text retrieval system the retrieval of text based information becomes much easier and user friendly, since one can use natural languages techniques such as automatic spell checking and stemming. The free text retrieval system needs first to index the database but then it is just to search the database.

Normally a search engine does not give any answers to queries when the search words does not exist in the index, therefore we connected a spell checker module into a search engine and evaluated it.

The domain used was the web site of the Swedish National Tax Board (Riksskatteverket, RSV), where the search engine was used between April and Sept 2001. One million queries were made by the public. Of these queries 10 percent were "misspelled" or erroneous and our spell checker corrected around 90 percent of these.

1 Introduction

A search engine is a device that reads through a document collection and indexes each word in each text and then builds an inverted index. The inverted index contains all the words in the text collection and each word points at its corresponding document/s. A search engine then searches the index to find the relevant documents. There is a number of different ranking models, e.g., Boolean ranking, term weight frequency and the vector space models, (van Rijsbergen, 1979).

Using a search engine entering key words does sometime not give any answer either because the word is not in the index or because the user misspelled the word, or because the user did not know the right inflection of the word as written in the index.

The information in a database is usually found using SQL or some other query language, but if one uses a free text retrieval system the retrieval of text based information becomes much easier and user friendly. The free text retrieval system needs first to index the database but then it is just to search the database.

2 Previous research

One method to increase recall in information retrieval is to use truncation or manual word expansion but the most correct and efficient method is automatic inflection or stemming

An automatic stemmer for Swedish is described in (Carlberger et al 2000), where it is shown that for Swedish, stemming improves the precision and recall by 15 percent and 18 percent

respectively¹⁾. One of the first attempts to use something between stemming and spell checking to increase the recall was described by Pearce & Nicholas (1993). They called it similarity score. One description of a spell checker for English in a search engine can be found in Hodge & Austin (2001), where recall is calculated to 97.5 percent. Both search engines Google (www.google.com) and AltaVista (www.altavista.com) have spell checking systems connected to the search engines though no evaluation is available. (We asked them and they said that their spellchecker was good, but not how good in figures).

Misspelling can be carried out by three reasons either because the user does not know the spelling or because of a typing error or because the user is not completely sure about the spelling

Sometimes the mistake can be different spelling as: *Kyrkskatt* / *Kyrkoskatt* (Church tax). This is similar to British or American spelling or that the concept can be close but not really right (See Appendix).

According to Knutsson (Knutsson 2001) there are around 2.4 percent spelling errors in texts written by second language users of Swedish.

In an investigation made by Hultman & Westman, (1977) on handwritten essays made by Swedish high school students they had a range from 0.4 percent spelling errors for the best students to 1.2 percent spelling errors for the worst students. According to Kukich (1992) there are around 0.2 percent to 3 percent spelling errors in English texts.

3 Issues

- Swedish word are very often compounds. What happens if the spell checker splits the word in two or more elements, do the user obtain more answers? Does recall increase?
- What is a misspelled word in the document collection?
- If the word occurs at least twice in a small document collection (< 6000 document) is it then a correct word?
- How do we treat "misspelled" words in the document collection?
- Should misspelled words be indexed?
- Are the proposed corrections valid or valuable?
- Should the search engine propose not "correct" spelled word that can be in the retrieved text?
- Can bad suggestions be filtered away?
- One other very important issue is how much a spelling correcting algorithm for Swedish (or any other language) connected to a search engine would improve the precision and recall?

Search scenarios

1) *Search and find*

This is the normal *successful* case

¹⁾ Precision = number of found relevant documents / total number of found documents
Recall = number of found relevant documents / total number of relevant documents

- 2) Search and not find and no proposal of other spelling
This is the other normal *not successful* case in a “normal” search engine.
- 3) *Search and find and propose other spellings simultaneously.*
This is possible the best scenario, the user obtains answer and gets also some possible alternative spellings.
- 4) *Search and not find and propose other spellings.*
This is the RSV case where one gets some spelling proposals if no answer is found.
- 5) *The proposed spelling can be spelling mistakes performed in the text collections by an author.*
This scenario can give wrong impression to the seeker, how should we treat these examples? Maybe scenario 3) can solve this problem. One more possible option is to tell the user the number of hits with this particular spelling so that s/he can get a picture of the frequency of the spellings.
- 6) *The spell checker splits compounds in two or more words.*
This is important since Swedes are very often influenced by English compound splitting of words in their Swedish writing.
- 7) *Spell check of a search word performing dictionary lookup*
Not a good idea since the proposed spelling might not be at all in the index and using the index as a will also make it possible to make spelling corrections in many languages not only Swedish.

4 Building a spelling correction algorithm

According to Kukich (1992), the four error types, insertion, deletion, substitution and transposition encompasses 80 percent of all spelling errors, therefore we think that connecting a spell checker with a search engine will assist the searcher a lot.

Our spell checking algorithm stems from Stava and Granska (Domeij et al 1994, Carlberger & Kann 1999, 2000, Knutsson 2001) and makes specifically use of the Edit-distance techniques described in (Kukich 1992). Normally spell checkers use string matching techniques to a specific dictionary. In a search engine the spell checker uses the index as lexicon, otherwise it might propose words which are not in the index.

5 Experiment

From April to September 2001 the Swedish National Tax Board (Riksskatteverket, RSV) used Euroseek’s search engine Euroseek Remote Indexing with Eurolings (KTH) built-in stemmer and dynamic spell checker.

The RSV site had almost 6000 documents mostly in Swedish containing almost 11 million words. There is a very large discrepancy in the size of each document ranging from very small documents around a half a page to several megabytes.

During the testing period from April to September 2001 we obtained 1 031 700 queries to the search engine of these queries 101 446 (9.8 %) where errors (spelling errors, similar spelling or words not in the index) to which our algorithms proposed alternative words, of these 9 percent where bad alternatives and to 36 253 (3.5%) of the total queries the system could not give any alternative at all (not in the Appendix).

Of the top 100 spelling errors the system gave 92 percent good suggestions and 40 percent of these contained split compound words, 22 percent was spelling errors and 30 percent was alternative spellings.

Some examples: (word => good suggestion)

40 percent of the good suggestions were compound splitting, see examples below.

utrikestraktamente	=>	traktamente utrikes (allowance abroad)
bilavgifter	=>	avgifter bilar
experts katt	=>	expert skatt
skattejämningsblankett	=>	jämningsblankett skattejämnning

22 percent of the good suggestions were plain spelling errors, see examples below.

engångskatt	=>	engångsskatt
gitemål	=>	giftmål (marriage)
jämnkning	=>	jämnkning
skilsmässa	=>	skilsmässa (divorce)
skiljsmässa	=>	skilsmässa
skattejämnkning	=>	skattejämnkning

30 percent of the good suggestions were alternative spellings or stemming, see examples below.

engångskatt	=>	engångsskatt
kyrskatt	=>	kyrkoskatt (church tax)
hempc	=>	hem-pc
rotavdrag	=>	rot-avdrag
arvsskifte	=>	arvsskiftet
pharmasia	=>	pharmacia
skatteåterbäring	=>	skatteåterbäring

Regarding compound splitting there is good idea to distinguish two cases, either there is *traktamente* OR *utrikes* OR there is *traktamente* NEAR *utrikes*, the latter example means that the words go together and have some relation.

In the Appendix we can see parts of the logs from RSV, we can also see that from the logs one can build or hard code a synonym list to directly help the user. For example if a common spelling error does not give any result then one can hard code the correct result.

Yes, one should propose "misspelled" words when found in index. The automatic indexing system does not know if a word is misspelled or not, since the system does not have a lexicon to compare with. In the RSV case there is the word *fastighetsskatt* meaning *tax on real estate* which is misspelled to *fastighetskatt*, (with one "s"), meaning *cat on real estate*. That misspelled

document contained important information on *tax on real estate* that of course should be found by the search engine and presented.

6 Conclusion and future improvements

We found that 10 percent of all queries to a search engine were in some sense erroneous or where not present in the index. Using our spelling correcting algorithm corrected 92 percent of these errors, 40 percent of all suggestion included compound splitting to obtain hits in the document collection, 22 percent were spelling errors and 30 percent were similar spelling or stemming.

To reveal the precision and the recall of the spell checker we would propose a new experiment similar to the stemming experiment performed in Carlberger et al (2001), where we compared precision and recall with and without a stemmer. The new experiment would be to compare precision and recall with and without a spell checker.

The experiment though need to be performed by giving the user implicit written queries so s/he can not see the spelling or give them correct keywords orally.

This paper presents a novel approach and needs to be further evaluated.

Acknowledgement

I would like to thank Johan Carlberger at Euroling AB for the implementation and integration of the spelling correction algorithm with the search engine. I would also like to thank Ola Knutsson, Richard Domeij and Martin Hassel at NADA-KTH for valuable comments on the paper and pointers to the literature.

References

- Carlberger, J., H. Dalianis, M.Hassel and O. Knutsson. Improving Precision in Information Retrieval for Swedish using Stemming. In the Proceedings of NODALIDA '01 - 13th Nordic Conference on Computational Linguistics, May 21-22, 2001, Uppsala, Sweden, 2001.
- Carlberger, J. and V. Kann. 2000. Some applications of a statistical tagger for Swedish. Proc. 4:th conference of the International Quantitative Linguistics Association (Qualico-2000), pp. 51-52, August 2000
- Carlberger, J. and V. Kann. Implementing an efficient part-of-speech tagger. Software Practice and Experience, 29, pp. 815-832, 1999.
- Domeij, R., J.. Hollman, and Viggo Kann. Detection of spelling errors in Swedish not using a word list en clair. Journal of Quantitative Linguistics 1:195-201, 1994.
- Hodge, V and J. Austin. A Comparison of a Novel Neural Spell Checker and Standard Spell Checking Algorithms, accepted for Pattern Recognition, Elsevier Publications, 2001

Hultman, T. and M, Westman,. Gymnasistsvenska, LiberLäromedel, Lund, 1977.

Knutsson, O. Automatisk språkgranskning av svensk text (in Swedish), (Automatic Proofreading of Swedish text), Licentiate Thesis. IPLAB-NADA, Royal Institute of Technology, KTH, Stockholm, 2001.

Kukich, K. Techniques for automatically correcting words in text, ACM Computing Surveys. Vol.24, No. 4 (Dec. 1992), pp. 377-439, 1992.

Pearce, C and C. Nicholas. Generating a Dynamic Hypertext environment with N-gram Analysis. Proceedings of the second International Conference on Information and knowledge Management, Washington DC, USA, pp. 148-153, 1993.

van Rijsbergen, C. J. Information retrieval. Butterworths, 1979

Appendix: 100 most common spelling errors

100	most common spelling errors	Spelling suggestion No 1	No 2	No 3	No 4					
No	of spelling errors	Objective ranking	Objective ranking	Objective ranking	Objective ranking					
1204	taxeringskalender	1	taxeringsr	0	taxeringsnämnder	1	taxering	1	taxeringsärendet	1
829	akassa	1	a-kassa	1	kassa	1				
826	fackavgift	1	avgift	1						
338	hempc	1	hem-pc	1	hempå	0				
313	taxeringskalender	1	taxeringsr	0	taxering1	0	taxerings	1	taxeringsärendet	1
311	jämknning	1	jämknning	1						
290	allmänhetenstermin	1	allmänheten	1						
243	rsv2252		rsv260	1	rsv2341	1	rsv297	0	rsv2	0
195	taxeringskalendern	1	taxeringsr	0	taxeringsärenden	1	taxerings	0	taxering1	0
182	taxeringskalender		taxeringsärendet	1	taxeringsnämnder	1	taxeringsärenden	1	taxeringsärendet	1
174	personbeskattningsenheten	1	personbeskattning	1	persona	0	persone	0	persons	1
173	rsv379		rsv297	1	rsv354	1	rsv360	1	rsv409	1
167	vägs katt	1	skatt	1	värnskatt	1	vägskäl	0	vägrat	0
159	skillsmässa	1	skillsmässa	1						
150	rotavdrag	1	rot-avdrag	1						
141	skattejämknning	1	skattejämknning	1						
140	fackavgifter	1	fallavgifter	0	avgifter	1				
137	trygghansa		trygg-hansa	1						
134	kommunalskatt	1	kommunalskatt	1						
132	personaldator		persondator	1	personas	0				
131	giftermål	1	giftermål	1						
130	engångsskatttabell	1	engångsskatter	1	engångsskatt	1	skatttabell	1		
126	bouppteckning	1	bouppteckning	1						
121	rsv418		rsv408	1						
121	hinderprövning	1	hindersprövning	1						
121	epsilon	0	epiros	0						
115	personaldatorer	1	personatorer	1	personal	1				
110	rsv285		rsv297	1	rsv260	1	rsv2	1	rsv08a	1
110	kyrskatt		kyrskatt	1						
109	pergo	0	perso	0						
108	transaktionskod	1	transaktion	1	transaktionskostnad	1	transaktionslag	1	transaktionskedja	1
107	a-kasseavdrag	1	a-kasseavgift	1	a-skatteavdrag	1	a-kassa	1	avdrag	1
105	engångsskatt	1	engångsskatt	1						

56	bokmoms		0	bokförs	0	bokens	0	bomhus	0	bokform	0
55	överskjutandes katt		1	överskjutande	1	överskjutan	0	översko	0	överst	0
55	måltidsavdrag		1	måltid	1	avdrag	1				
54	rsv349			rsv354	1	rsv409	1	rsv330	1	rsv360	1
53	ludvikamoms		1	ludvika	1	ludvig	0				
52	förmånsvärdesberäkning		1	förmånsvärden	1	förmånsvärdet	1	förmånsvärde	1	beräkning	1
51	skattåterbäring	1		skatteåterbäring	1						
51	rsv355			rsv354	1						
12577	Top 100 spelling errors	22	48	Good suggestions	82	Good suggestions	43	Good suggestions	21	Good suggestions	19
			8	Bad suggestions	18	Bad suggestions	16	Bad suggestions	16	Bad suggestions	15

40 No of compound splittings 6835 (of total top 100)
22 No of spelling correctings 2325 (of total top 100)
30 No of similar spellings, similar words, other inflections, etc
8 No of bad suggestions

100 Of the top spelling errors

Total Good suggestions	165	61%	Percentage Good suggestions when at least one good suggestion
Total Bad suggestions	65		
Total queries		1031700	
Total spelling errors		101446	
God suggestions		92%	(calculated on top 100 spelling errors)