Managing Keyword Variation with Frequency Based Generation of Word Forms in IR

Kimmo Kettunen
Department of Information Studies
University of Tampere
kimmo.kettunen@uta.fi

Abstract

This paper presents a new management method for morphological variation of keywords. The method is called FCG, Frequent Case Generation. It is based on the skewed distributions of word forms in natural languages and is suitable for languages that have either fair amount of morphological variation or are morphologically very rich. The proposed method has been evaluated so far with four languages, Finnish, Swedish, German and Russian, which show varying degrees of morphological complexity.

1 Introduction

Word form normalization through lemmatization or stemming is a standard procedure in information retrieval because morphological variation needs to be accounted for and several languages are morphologically non-trivial. Lemmatization is effective but often requires expensive resources. Stemming is also effective, generally almost as good as lemmatization and typically much less expensive; besides it also has a query expansion effect. However, in both approaches the idea is to turn many inflectional word forms to a single lemma or stem both in the database index and in queries. This means extra effort in creation of database indexes.

In this paper we take an opposite approach: we leave the database index un-normalized and enrich the queries to cover for surface form variation of keywords. A potential penalty of the approach would be long queries and slow processing. However, we show that it only matters to cover a negligible number of possible surface forms even in morphologically complex languages to arrive at a performance that is almost as good as that delivered by stemming or lemmatization. Moreover, we show that, at least for typical test collections, it only matters to cover nouns and adjectives in queries. Furthermore, we show that our findings are particularly good for short queries that resemble normal searches of web users.

Our approach is called FCG (for Frequent Case (form) Generation). It can be relatively easily implemented for Latin/Greek/Cyrillic alphabet languages by examining their (typically very skewed) nominal form statistics in a small text sample and by creating surface form generators for the 3−9 most frequent forms. We demonstrate the potential of our FCG approach for four languages of varying morphological complexity: Swedish, German, Russian, and Finnish in well-known test collections (CLEF 2003 and 2004). Applications include in particular Web IR in languages poor in morphological resources.

2 Word Form Distributions

It is well known that the distributions of words and word forms are not even in texts. Some word forms occur often, some are rare. Even the distributions of different morphological categories have rates of their own, and both semantic and morphological factors play a role in distribution of word form frequencies (Baayen, 1993, 2001). Karlsson (1986, 2000) shows with some semantically distinctive word types, how the case distributions of the words differ in Finnish. A word denoting a place, like

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 318–323

Helsinki, has besides the dominating nominative and genitive singular forms mainly occurrences of locative cases. A person's name like *Martti* occurs mostly in nominative singular. Same kind of analysis is given by Kostić et al. (2003) for Serbian, although they seem to be hesitant about the semantic origins of the phenomenon. We shall not explore the semantic factors of case distribution any deeper, but analyze the distribution of cases on morphological level only.

In Kettunen and Airio (2006) we first sought for corpus statistics of Finnish nominal word forms. Then we verified these statistics with two independent automatic analyses of larger corpuses. Our analysis and earlier corpus statistics showed, that six cases (out of 14) constituted about 84 – 88 % of the token level occurrences of case forms for nouns – thus covering 84 – 88 % of the possible variation of about 2000 distinct inflectional forms of nouns. Our analysis also showed that the huge number of grammatical forms is mainly due to clitics and possessive endings that are almost nonexistent even in a reasonably large textual corpus (10.3 M nouns). This analysis demonstrated that, while a language may in principle be morphologically complex, in practice it is much less so.

2.1 Distribution Based Management of Keyword Variation for IR

Our FCG method and its language specific IR evaluation are simply as follows:

1) For a morphologically complex enough language the distribution of different nominal case/other word forms is first studied through corpus analysis (if such results are not available for the language). The used corpus can be quite small, because variation at this level of language can be detected even from smaller corpuses. Variation in textual styles may affect slightly the results, so a style neutral corpus is the best. If style specific results are sought for, then an appropriate corpus needs to be used in word form occurrence analysis.

2) After the most frequent (case) forms for the language have been found with corpus statistics, the IR results of using only these forms for noun and adjective keyword forms are evaluated in a standard IR collection. As a comparison best available normalization method (lemmatization or stemming) is used. The number of tested FCG processes depends on the morphological complexity of the language: more processes can be tested for a complex language, only a few for a simpler one.

3) After evaluation, the best FCG process with respect to normalization is usually distinguished. The evaluation process will probably also show that more than one FCG process is giving quite good results, and thus a varying number of keyword forms can be used for different retrieval purposes, if necessary.

We have been simulating the process of keyword generation in our tests, but as word form generation programs are available for many languages, their output could be modified accordingly for real use, i.e., only the most frequent forms of generated forms would be used in search.

Based on this method, we evaluated four different FCGs in two different full-text collections of Finnish, TUTK (with multi-valued relevance) and CLEF 2003 (with binary relevance) with long title and description queries. The results of Kettunen and Airio (2006) showed that frequent case form generation works in full-text retrieval of inflected indexes in a best-match query system and competes at best well with the gold standard, lemmatization, for Finnish. Our best FCG procedures, FCG_9 and FCG_12 - with 9 and 12 variant keyword forms - achieved about 86 % of the best average precisions of FINTWOL lemmatizer in TUTK and about 90 % in CLEF 2003. We thus performed successful information retrieval of Finnish with nine and twelve variant keyword forms, which is 0.48 % and 0.64 % of the possible grammatical forms of Finnish nouns ($\sum$ = 1872) and about 34.6 % and 46.2 % of the productive forms ($\sum$ = 26).

We now evaluated performance of Finnish short title queries in the CLEF 2003 collection. Results of the Finnish short queries (mean length 2,55 words when stop words were omitted) are shown in Table 1.

Method	Mean average precision
FINTWOL, compounds split	42.8 %
Stemmed	41.3 % (-1.5)
FINTWOL, compounds not split	40.5 % (-2.3)
FCG_12	38.1 % (-4.7)
FCG_9	37.9 % (-4.9)
Inflected	22.6 % (-20.2)

Table 1. Finnish CLEF 2003 results, 45 title queries

As can be seen from Table 1, difference between the best FCG method and the best achieved results, FINTWOL with index where compounds are split, is about 5 absolute per cent with short queries. Thus the method works also well with short and realistic queries, and about 88−89 % of the maximal retrieval result is achieved with both nine and twelve most frequent nominal forms of the keywords.

Figure 1 shows P/R-curves of the best Finnish FCG procedure (FCG_12), FINTWOL with split compound index and plain query words for short queries.

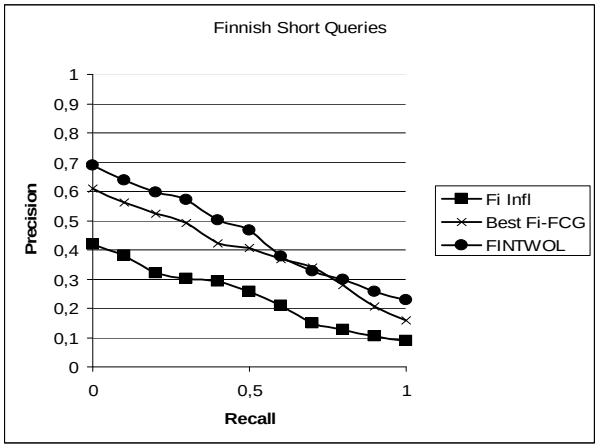

Figure 1. P/R-curves for Finnish short queries: precision by eleven recall levels 0.0−1.0

3 FCGs for Three More Languages

In this study we evaluated further our word form frequency based method with three European languages, Swedish, German, and Russian. They are all morphologically moderately complex, i.e. clearly much more complex than English, but also clearly much simpler than Finnish (or Hungarian) measured in the number of possible word forms per lexeme. The chosen languages represent two major language groups of the Indo-European language family, Germanic (German and Swedish) and Slavonic (Russian), and are thus also characteristic samples for other languages in the same language groups. The languages were chosen on the basis of available IR collections and complex enough nominal morphology from the CLEF materials. From the morphological complexity point of view there would have been other and perhaps more interesting languages among the official EU languages (e.g., Estonian, Lithuanian, Latvian, Slovak, Czech and Hungarian), but either lack of available IR collections or detailed enough linguistic knowledge in the languages made inclusion of these languages impossible in this study.

3.1 Materials and Methods

CLEF collections for all the three languages were utilized in this study. For Swedish and German we used materials of CLEF 2003. The retrieval system was InQuery. For Russian we used Russian collection of CLEF 2004 and the Lemur retrieval system. Character encoding for Russian was UTF-8. In Table 2, the number of documents and topics in each collection is shown (Airio, 2006; Tomlinson, 2004).

Language	Collection	Collection size (docs)	Topics
Sv	CLEF 2003	142 819	54
De	CLEF 2003	294 809	56
Ru	CLEF 2004	16 716	34

Table 2. Swedish, German and Russian collections used in the study

For Swedish we analyzed the most frequent word forms to be used as keywords in FCG queries on the basis of a SWETWOL analysis of newspaper material from Helsingborgs Dagblad 1994 and Göteborgs posten 1994 texts, altogether 161 336 articles (Ahlgren, 2004, 61). For German word form frequency analysis we used an existing

morphologically annotated Tiger corpus. For Russian we obtained the case distribution information from the Russian national corpus. Statistics of the distribution analysis are published in Kettunen et al. (2007).

On the basis of these corpus analyses we formed FCG procedures for each language with different number of keywords. Swedish got two procedures, Sv-FCG_2 and Sv-FCG_4, as did also German with procedures De-FCG_2 and De-FCG_4. As Russian was morphologically the most complex language of these, it got three FGC procedures, Ru-FCG_3, Ru-FCG_6 and Ru-FCG_8. The figure in the name of the procedure gives the approximate number of morphological keyword variants for each procedure.

Queries for the FCG procedures of each language were formed manually from the topics by using different language tools in the web (electronic dictionaries, word form generators or both). After we had formed the queries, we evaluated the retrieval results for each language. As a comparison we used lemmatization with TWOL programs for Swedish and German and also Snowball stemmers for both of the languages. For Russian we only had access to Snowball stemmer. Also plain topic words were used in the queries of all languages to get a baseline result.

Our queries were structured with InQuery's #SYN operator. With the operator morphological variant forms of the keyword are treated as synonyms of the key, and InQuery treats them all as instances of one key.

As a FCG query example we can take one query from the CLEF 2003 collection. A short version of query #142 for the Sv-FCG_4 process is as follows:

```
#q142 = #sum(#syn( christo )
#syn( paketerar ) #syn( det )
#syn( tyska tyskt ) #syn(
riksdagshuset riksdagshus
riksdagshusen));
```

As can be seen from the query example, only nouns and adjectives of the query are expanded with variant forms, all words of other categories are left in the form they were in the original topic. Nouns are self-evidently most important for queries, but adding variant forms of adjectives seems also to increase mean average precision of queries with 1−3 % in each language.

4 Results

4.1 Swedish results

We ran both long and short queries for all the languages. Here we show and discuss only results of short title queries. Full results are presented in Kettunen et al. (2007).

Results of the Swedish very short queries (average length 3.17 words with stop words) are shown in Table 3.

Method	Mean average precision
SWETWOL, compounds split	32.6 %
Sv-FCG_4	30.6 % (-2.0)
Sv-FCG_2	29.1 % (-3.5)
Stemmed	28.5 % (-4.1)
SWETWOL, compounds not split	26.3 % (-6.3)
Inflected	24.0 % (-8.6)

Table 3. Results of the 54 Swedish title queries

SWETWOL with split compounds in the database index gets the best results, but the best Sv-FCG procedure is not far behind. The margin between non-processed keywords and best normalization result is 8.6 %. Both Sv-FCGs outperform stemming and SWETWOL without compound splitting.

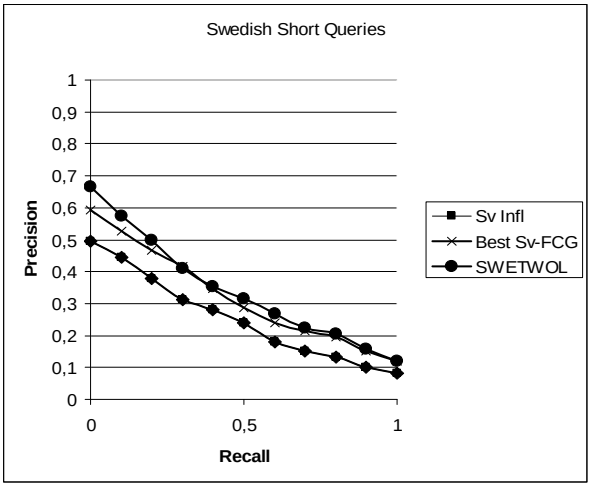

Figure 2. P/R-curves for Swedish short queries: precision by eleven recall levels 0.0−1.0

Figure 2 shows P/R-curves of the best Swedish FCG procedure (Sv-FCG_4), SWETWOL with split compounds and plain query words for short queries.

4.2 German results

Results of the German very short queries (average length 3.15 words with stop words) are shown in Table 4.

Method	Mean average
GerTWOL, compounds split	29.6 %
Stemmed	30.9 % (+1.3)
De-FCG_4	29.9 % (+0.3)
De-FCG_2	29.0 % (-0.6)
GerTWOL, compounds not split	28.1 % (-1.5)
Inflected	25.4 % (-4.2)

Table 4. Results of the 56 German title queries

The Snowball stemmer performs the best with a 1.3 % margin to GERTWOL using split compound index. De-FCG_4 is also slightly better than GERTWOL, and De-FCG_2 outperforms also GERTWOL without compound splitting. Non-processed queries perform worst, and the margin of non-processing to the best performing system, Snowball, is 5.5 %. The margin of non-processing to the worst performing normalization is 2.7 %.

Figure 3 shows P/R-curves of the best German FCG procedure (De-FCG_4), German Snowball and plain query words for short queries.

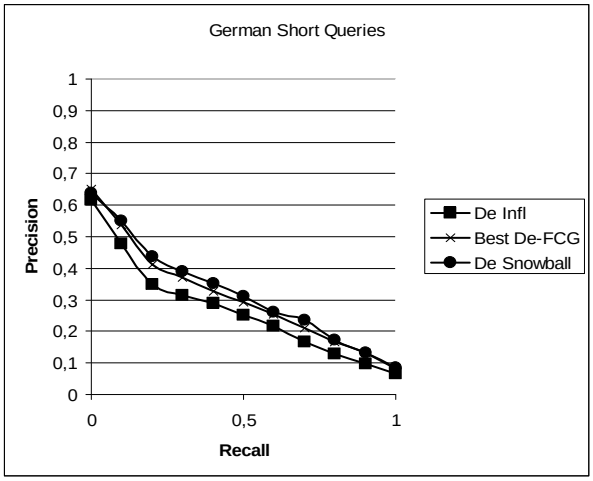

Figure 3. P/R-curves for German short queries: precision by eleven recall levels 0.0−1.0

4.3 Russian results

Results for Russian short queries are shown in Table 5. Mean length of the queries was 3.18 words (with stopwords).

Method	Mean average precision
Ru-FCG_6	32.0 %
Ru-FCG_8	31.7 % (-0.3)
Ru-FCG_3	31.2 % (-0.8)
Snowball Ru	27.2 % (-4.8)
Inflected	25.1 % (-6.9)

Table 5. Results of 34 Russian title queries

Our Russian results are not as clear as those of Swedish and German, because results of long and short queries in Russian were quite different. Overall it seems that short Russian queries show some advantage for FCGs, but as the collection is small and has very few relevant documents, the interpretation of the Russian results remains inconclusive.

Figure 4 shows P-R-curves of the best Russian FCG procedure (Ru-FCG_6), Russian Snowball and plain query words for short queries.

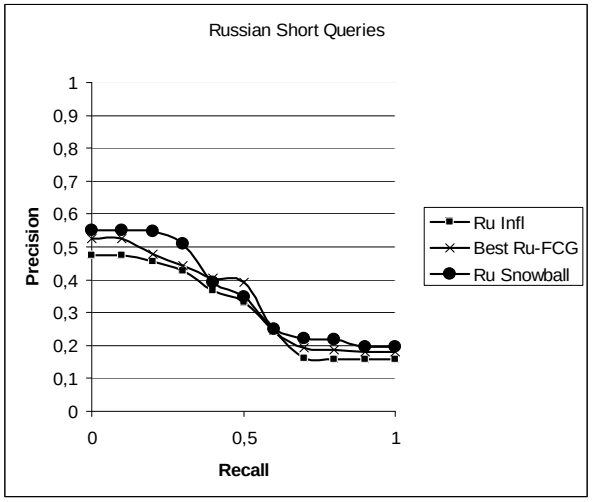

Figure 4. P/R-curves for Russian short queries: precision by eleven recall levels 0.0−1.0.

5 Discussion

The main reason for using stemming, lemmatization or any kind of morphological processing with IR is improvement in precision and recall of searches. Although the gains of morphological processing are varying, they are real. The usual

way to estimate the performance gains is relative percentage improvement of mean average precisions between different methods. For comparison purposes of methods a slightly different point of view could also be used: the difference between doing nothing for the query words and the best mean average precision shows the need of morphological processing for the language in question. The bigger the discrepancy between these figures, the bigger the need to do something for the keywords.

In Figures 1−4 P/R-curves of Finnish, Swedish, German and Russian short queries for the best normalization method, best FCG method and no processing at all were shown. As can be seen from the figures and Tables 1, 3, 4 and 5, the largest difference between non-processing and best normalization method is in Finnish (20.4 %) and smallest in Swedish (4.1 %). German and Russian have slightly greater differences than Swedish, 5.7 % and 6.9 % respectively. Figures show that the FCG method gives clear gains for Finnish and smaller gains for German, Swedish and Russian. For three languages FCG works well in comparison to lemmatization; for Finnish 88 % of the performance of lemmatization is achieved and 95 % for Swedish and German. The P/R graphs also show that the FCG method pushes close to normalization even when the gap between normalization and non-processing is narrow. Gains over no morphological processing at all are greater than losses against normalization.

For the three new languages evaluated two, Swedish and German, showed quite clearly that the FCG method works well for both languages. In short queries the differences between all the methods were smallest, but also the margin between plain keywords and the best method increased. In German runs overlap of inflectional noun forms slightly disturbed results.

Our Russian results remained partly counterintuitive. Although recall rose steadily when more case forms were put into the query, the mean average precision of short queries did not get much better, when forms were added. Overall it seemed that short Russian queries showed some advantage for FCGs. As the collection was small and had very few relevant documents, the interpretation of the Russian results remained inconclusive. Thus the method should be re-evaluated in a better Russian collection.

References

Per Ahlgren. 2004. *The effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database*. Department of Library and Information Science/Swedish School of Library and Information Science. University college of Borås/Göteborg University.

Eija Airio. 2006. Word Normalization and decompounding in mono- and bilingual IR. *Information Retrieval 9*: 249–271.

R. Harald Baayen. 1993. Statistical Models for Word Frequency Distribution. *Computers and the Humanities* 26: 347–363.

R. Harald Baayen. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht Boston London.

Fred Karlsson. 1986. Frequency Considerations in Morphology. *Zeitsschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 39: 19–28.

Fred Karlsson. 2000. Defectivity. In: Booij G. et al. (eds.): *Morphology. An International Handbook on Inflection and Word-Formation.* Volume 1. Walter de Gruyter, Berlin, 647–654.

Kimmo Kettunen and Eija Airio. 2006. Is a morphologically complex language really that complex in full-text retrieval? In T. Salakoski et al. (Eds.): *Advances in Natural Language Processing,* LNAI 4139. Springer-Verlag Berlin Heidelberg, 411–422.

Kimmo Kettunen, Eija Airio and Kalervo Järvelin. 2007. Restricted Inflectional Form Generation in Management of Morphological Keyword Variation. *Information Retrieval* (to appear).

Alexandar Kostić, Tanja Marković and Alexandar Baucal. 2003. A. Inflectional Morphology and Word Meaning: Orthogonal or Co-implicative Cognitive Domains. In: Baayen, R.H. and Schreuder R. (eds.): *Morphological Structure in Language Processing*. Trends in Linguistics, Studies and Monographs 151. Mouton de Gruyter, Berlin, 1–43.

Tiger corpus. http://www.ims.uni-stuttgart.de/projekte/-TIGER/TIGERCorpus/ (visited June 7th, 2006).

Stephen Tomlinson. 2004. Finnish, Portuguese and Russian Retrieval with Hummingbird SearchServerTM at CLEF 2004. Working Notes for the CLEF 2004 Workshop, 15-17 September, Bath, UK. <http://clef.isti.cnr.it/2004/working_notes/WorkingNotes2004/21.pdf>. Accessed October 10th, 2006.