

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

Kimmo Kettunen

Dept. of Information Studies,
University of Tampere, Finland
FIN-33014 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.

Categories and Subject Descriptors

H.3.3 Information Search and Retrieval H.3.1 Content Analysis and Indexing *Linguistic processing*

General Terms

Measurement, Performance, Design, Experimentation, Languages.

Keywords

Management of morphological variation, monolingual information retrieval, evaluation, word form generation

1. INTRODUCTION

Word form normalization with lemmatization or stemming is a standard procedure in information retrieval because morphological variation needs to be accounted for and many 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. In both approaches the idea is to turn many inflected word forms to a single lemma or stem both in the database index and in queries. This means extra effort in creating 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 either lemmatization or stemming. We also show that, at least for typical test collections, it only matters to cover variation of nouns and adjectives in queries; cf. for further argumentation in [6]. 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. DISTRIBUTIONS OF WORD FORMS

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 [1, 2]. Karlsson [3, 4] shows with some semantically distinctive word types, how the case distributions of the words differ in Finnish. Same sort of analysis is given by Kostić et al. [7] for Serbian and by Perebeynoss and Khikedel [8] for English verb forms. We shall not explore the semantic factors of case distribution any deeper, but analyze the distribution of cases on morphological level only.

We introduced our word form distribution based FCG method in [5] with Finnish. First we sought for corpus statistics of Finnish nominal word forms and then 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 inflected forms of nouns.

2.1 Distribution Based Handling of Keyword Variation for IR

Our FCG method and its language specific evaluation procedure 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. 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.

2) After the most frequent (case) forms for the language have been identified with corpus statistics, the IR results of using only these forms for noun and adjective keyword forms are tested in a well-known test collection. As a comparison best available keyword and index normalization method (lemmatization or stemming) is used, if such is available. 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 testing 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.

Based on this method, we first evaluated four different FCGs in two different full-text collections of Finnish, TUTK (with multi-valued relevance) and CLEF 2003 (with binary relevance). The results of [5] 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 for nouns and adjectives - achieved about 86 % of the best average precisions of FINTWOL lemmatizer in TUTK and about 90 % in CLEF 2003.

2.2 FCGs for Three More Languages

In this study we evaluated further our word form frequency based method with three more 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 nominal lexeme. The languages were chosen on the basis of available IR collections and complex enough nominal morphology from the CLEF materials.

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 generated word forms would be used in search.

3. RESULTS

Due to lack of space we only present P/R-curves of Swedish and German short queries. More detailed results will appear in Kettunen, Airio and Järvelin [6]. Figure 1 shows P/R-curves of the best Swedish FCG procedure (Sv-FCG_4), SWETWOL lemmatizer and plain query words for short queries made out of the titles of the topics.

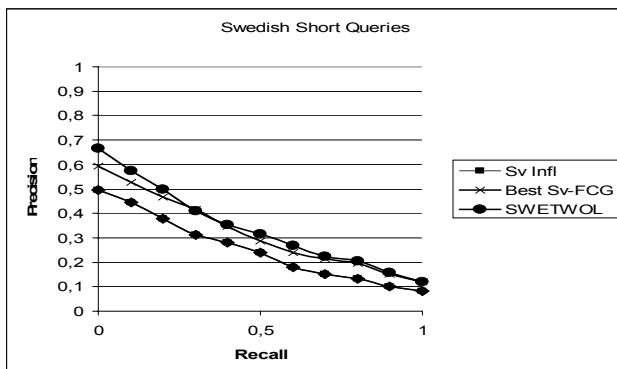


Figure 1. P/R curves for Swedish short queries: precision by eleven recall levels 0.0 - 1.0

Figure 2 shows P/R-curves of the best German FCG procedure (De-FCG_4), German Snowball and plain query words for short queries.

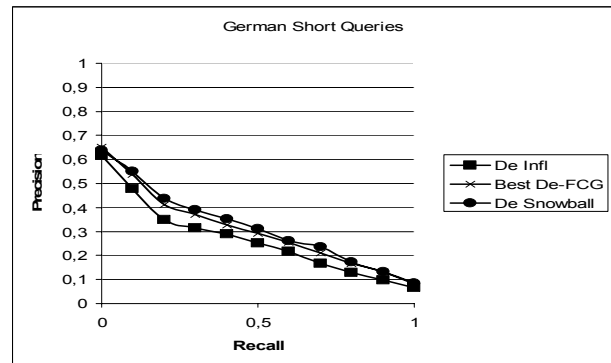


Figure 2. P/R-curves for German short queries: precision by eleven recall levels 0.0–1.0

Our Swedish and German results showed quite clearly that the FCG method works for Swedish and German both in long and short queries. In short queries differences between all the methods were smallest in German tests the overlap of inflected noun forms slightly disturbed results. Our Russian results remained partly counterintuitive. With both long and very short queries recall rose steadily when more case forms were put in to the query. Anyhow, the mean precision of long queries did not get any better, when forms were added, but rather decreased. Short Russian queries showed some advantage for FCGs, but as the collection was small and had very few relevant documents, the interpretation of the Russian results remained inconclusive.

4. REFERENCES

- [1] Baayen, R. H. Statistical Models for Word Frequency Distribution. *Computers and the Humanities* 26 (1993): 347–363.
- [2] Baayen, R. H. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht Boston London, 2001.
- [3] Karlsson, F. Frequency Considerations in Morphology. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 39 (1986): 19–28.
- [4] Karlsson, F. Defectivity. In: Booij G. et al. (eds.): *Morphology. An International Handbook on Inflection and Word-Formation*. Volume 1. Walter de Gruyter, Berlin, 2000, 647–654.
- [5] Kettunen, K. and Airio, E. 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, 2006, 411–422.
- [6] Kettunen, K., Airio, E. and Järvelin, K. Restricted Inflectional Form Generation in Management of Morphological Keyword Variation. *Information Retrieval* (to appear).
- [7] Kostić, A., Marković, T. and Baucal, 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, 2003, 1–43.
- [8] Perebeynoss, V. and Khidekel, S. Frequency of Language Units as a Reflection of Their Systemic and Functional Properties. *Journal of Quantitative Linguistics* 11 (2004): 3–25.