Experiments to investigate the connection between case distribution and topical relevance of search terms in an information retrieval setting

Jussi Karlgren[1], **Hercules Dalianis**[2,3], **Bart Jongejan**[4]

[1]SICS, Box 1263, 164 29 Kista, Sweden
[2]DSV, KTH - Stockholm University, Forum 100, 164 40 Kista, Sweden
[3]Euroling AB, Igeldammsgatan 22c, 112 49 Stockholm, Sweden
[4]CST, University of Copenhagen, Njalsgade 80, 2300 København S, Denmark
jussi@sics.se, hercules@dsv.su.se, bart@cst.dk

Abstract

We have performed a set of experiments made to investigate the utility of morphological analysis to improve retrieval of documents written in languages with relatively large morphological variation in a practical commercial setting, using the SiteSeeker search system developed and marketed by Euroling AB. The objective of the experiments was to evaluate different lemmatisers and stemmers to determine which would be the most practical for the task at hand: highly interactive, relatively high precision web searches in commercial customer-oriented document collections. This paper gives an overview of some of the results for Finnish and German, and describes specifically one experiment designed to investigate the case distribution of nouns in a highly inflectional language (Finnish) and the topicality of the nouns in target texts. We find that topical nouns taken from queries are distributed differently over relevant and non-relevant documents depending on their grammatical case.

1. Morphological variation and its consequences for commercial provision of information retrieval systems

We have performed a set of experiments made to investigate the utility of morphological analysis to improve retrieval of documents written in languages with relatively large morphological variation in a practical commercial setting, using the SiteSeeker search system developed and marketed by Euroling AB[1]. The objective of the experiments was to evaluate different lemmatisers and stemmers to determine which would be the most practical for the task at hand: highly interactive, relatively high precision web searches in commercial customer-oriented document collections. [2]

This paper gives an overview of some of the results, and describes specifically one experiment designed to investigate the case distribution of nouns in a highly inflectional language (Finnish) and the topicality of the nouns in target texts. Nouns are chosen as the only lexical category to investigate, since we know from tracking our search logs that close to all terms in short web searches are nouns or proper names.

2. An evaluation of benefits of morphological analysis

2.1. Experimental setting

The annual Cross-language Evaluation Forum (CLEF) conferences provide a stable test bench of document collections, queries ("topics"), and manually obtained relevance judgements which relate sets of documents to topics. Each topic thus has a set of relevance judgments to select which documents are judged topically relevant to it. Typically the number of topically relevant documents for a topic is on the order of a few dozen and the number of assessed documents around two hundred.

Label	Query form	Corpus
oo	original	original
so	stemmed	original
mo	manual analysis	original
om	original	morph analysis
sm	stemmed	morph analysis
mm	manual analysis	morph analysis

Table 1: Evaluation cases

The document databases used for this set of experiments were the CLEF collection of Finnish newsprint and the SDA section of German-language newswire reports for the years 1994 and 1995; the seventy-five Finnish-language and one hundred German-language topics for the 2002 and 2003 evaluation cycles (numbers 91-200; title fields only; of the 110 topics only 75 and 100 had retrieved any relevant documents, respectively), and the relevance judgements given for those topics. The document collection consists of some 45 000 assessed documents of news articles, most rather brief, with about 1 000 relevant Finnish-language documents, and about 2 000 relevant German-language documents. Most queries consist of two to four words (Peters et al., 2003; Peters et al., 2004).

The text collections are processed morphologically, the Finnish corpus using the commercially available tools available from Connexor (Tapanainen and Järvinen, 1997); the German using a lemmatiser developed at CST (Jongejan and Haltrup, 2005). The hand crafted Euroling SiteSeeker stemmer has previously been compared with the CST lemmatizer that is developed using machine learning techniques, (Dalianis and Jongejan, 2006).

[1]http://www.euroling.se/en/siteseeker

[2]This work was partly supported by VINNOVA (the Swedish Governmental Agency for Innovation Systems) through the TvärSök project.

The evaluations are run in parallel using the analysed and the non-analysed corpora. There are two experimental cases for the corpus: o (original), using the unanalysed corpus with original word forms and m (morphological analysis) using the morphologically analysed corpus with lemmatised word forms. The SiteSeeker search system has for Finnish and German a built-in morphological normalisation component for queries, based on the widely available, well-established, simple, and robust Snowball-Porter stemmer (Porter, 1980). In the evaluations we used three experimental settings with regard to query processing: o (original) with no morphological analysis and the Porter stemmer disabled; s (stemmed), with Porter stemmer switched on; m (manual morphological analysis) with every query word manually taken to its lexical base form. This gives us six experimental cases in all for the evaluation, given in Table 1.

2.2. Results from the evaluation

A compilation of results from the evaluation are given in Table 2 and are given using standard precision and recall metrics as calculated by the trec_eval package[3], such as map (mean average precision), $bpref$ (an average precision measure which better models incomplete relevance judgements, (Buckley and Voorhees, 2004)), $P5$, and $P20$ (precision at rank 5 and rank 20, respectively). The results give us license to conclude that:

1. Careful – in this experiment, manual – normalisation of queries together with a morphologically analysed target corpus will by and large give the best results. Cases fi-mm and de-mm have best or near best results for most measures.

2. Even basic stemming improves results: so is better than oo and sm is better than om for both Finnish and German. The increase in precision at P5 for both Finnish and German is around 20 percent using the Snowball stemmer; recall shows a 133 percent increase for Finnish and a 94 percent increase for German, respectively. With morphological analysis of the target corpus the precision increase is over 40 percent.

3. If the documents are *not* morphologically analysed, inadvertent very careful morphological analysis of the queries *may* lower performance: in this experiment, case fi-mo has lower results than case fi-oo. But not necessarily: case de-mo is not bad at all. This reflects a typological difference between the languages: Finnish morphology is more expansive than is the German. This is presumably the explanation for case fi-mm not being better than case fi-sm. The manual morphological analysis has retrieved many further items, gaining considerably better recall, but losing precision.

4. The number of *returned* documents is a good indication of the number of *relevant* returned documents, which is a convenient proxy measure for inexpensive future evaluations.

These results conform with previous studies on expected and observed effects of using morphological analysis in information retrieval tasks. The results, especially the German-language ones, indicate that in spite of its simple construction and basic level of linguistic analysis, the Snowboll stemmer as used in SiteSeeker works better than expected. The Finnish results are less striking but still show an improvement in both recall and precision.

[3] Available from ftp://ftp.cs.cornell.edu/pub/smart.

3. Generative vs reductive approaches to handling morphological variation

The Finnish language still provides a challenge. In a recent series of publications, Kimmo Kettunen has experimentally shown how Finnish information retrieval tasks, hampered by the expansive morphology employed by the Finnish language, can be profitably performed by providing analyses informed by occurrence statistics rather than aiming for maximum coverage (2007). He shows in a series of experiments how *generative* methods, built to enhance the query by generating morphological variants of query terms, are more economic and nearly as efficient as *reductive* methods, building on normalising the index and query terms to a common normal representation, through stemming or lemmatisation.

3.1. Frequent case generation

For the Finnish language, where the expected gains of more informed morphological processing are likely to be great due to the expansive character of Finnish morphology[4], Kettunen finds that using the generally observed most frequent case forms for nouns – nominative, genitive, partitive in and the three inner locative cases[5] instead of trying for complete coverage of all possible forms of the noun gives a practical rendition of the actual usage, useful for information retrieval purposes. His experiments were performed by replacing each noun in the query by nine terms or twelve terms, depending on whether plural inner locatives are used or not (2006).

3.2. Comparison with best-case baseline

In Kettunen's and his colleagues' experiments, full, high-coverage, high-quality reductive index and query normalisation using TWOL, an established commercially available morphological tool based on two-level morphology (Koskenniemi, 1983), gives the consistently best results, as might be expected. However, their findings show that generative methods were often almost as useful as the best-case baseline and always better than doing no morphological analysis, raising recall noticeably. This was achieved at a much lower processing cost than doing full morphological processing.

[4] Finnish nouns can theoretically assume close to two thousand forms if all possible suffixes are factored in.

[5] Inner locative cases – inessive, illative, and elative – are cases to describe physical location with respect to something. They are prototypically used for expression of something being inside something, something entering or being inserted into something, or something exiting or being ejected from something else. They carry approximately the same meaning as English prepositional expressions with "inside", "in", "into", "out of": *pinteessä* "in trouble", *hississä* "inside the elevator", *ojasta allikkoon* "out of the frying pan into the fire", *Helsinkiin* "to Helsinki".

Kettunen and colleagues also found that in a strict relevance setting, where only the most relevant documents were targeted, the difference between the methods narrowed to less than in experiments where a more inclusive or liberal measure of relevance was used, indicating that in an interactive web service setting, where precision is the overriding concern, the savings occasioned by a generative method have less attendant cost than in a high-recall type scenario. In other experiments by the same team, similar results (although with somewhat lesser effect) are reported for Swedish and German (2007).

4. Does morphology mean anything?

Kettunen's point of departure, as is the case of most morphological analyses in information access experiments, is that morphological variation is noise and that conflation techniques should be introduced to reduce this noise as much as possible with least possible effort.

It is quite conceivable that morphology has as its sole function to organise local structure in the linguistic signal, and that once an utterance is perceived and understood the morphological variation given by its terms can be discarded as being irrelevant for text-level semantics. This is the view of most efforts in morphological analysis for information retrieval applications – many studies have been made for various languages showing how morphological normalisation of surface variation has beneficial effects on retrieval results.

However, it is also possible that the converse holds: the analysis of morphological variation may well yield semantic information on a level which can be utilised to better find relevant documents by its terms. This latter hypothesis, that of the text-level meaningfulness of morphological variation, has been tested in several experiments in the past, using information retrieval experimentation as a target, always with equivocal results for English or other languages with less elaborate morphologies.

A generative query processing framework allows hypothesis testing on another level. Terms can be treated heterogenously. Place names can be expanded in local cases, person names in animate cases, neither should be pluralised indiscriminately. Kettunen explicitly states he will not experiment using semantic distinctions, to avoid reliance on knowledge rich sources such as lexica, domain specific word lists, gazetteers, and the like. This is a reasonable research strategy, but in practical application these distinctions could be mined from the text using entity recognition algorithms.

Similarly, the case form distribution of terms in texts, irrespective of semantic qualities, could be mined to establish the topical poignancy of a term for a given topic. The topical qualities of a term in the target text could be expected to have influence on the morphological guise it is presented in. This should be easier to establish in languages with elaborate case marking, where e.g. some case forms typically indicate the head word is used as an adverbial rather than topical focus or centrality.

4.1. Query term frequencies in the target corpus

In the experiment presented here we take the search terms in the brief topic title queries used in the evaluation referred to in Section 2. and examine what case forms they occur in throughout the assessed relevant and non-relevant documents. We use the nine-form paradigm established by Kettunen to be most efficient: nominative, genitive, partitive in singular and plural and the three inner locative cases in singular.

For instance, for query 200 *Saksan ja Hollannin tulvat* ("Flooding in Holland and Germany"), we examine the case forms of *saksa* ("Germany"), *hollanti* ("Holland"), and *tulva* ("Flood") in the target texts. In Table 3 we show some observations of case forms of search words in the CLEF corpus for Finnish.

5. Observations

Our *methodological goal* in this given experiment is to build on hypotheses informed by some sense of textual reality, rather than computational expediency, and to evaluate the results by discriminatory potential of search terms. The alternative is to run an information retrieval task using some hypothesis as a basis for a new scheme to weight query and index terms: this a risky procedure if the goal is to understand the convergence of textual structure, terminological choice, and topical relevance since the effects may be completely overshadowed by noise or other, more powerful, topical factors.

While the full set of experiments aims at information retrieval effect, we have here studied the basis of word-occurrence based retrieval: occurrence of search terms in target texts. Without significant difference in occurrences of search terms in relevant vs non-relevant documents no effects in document retrieval can be expected to emerge from the full extrinsic information retrieval evaluation.

Our overriding hypothesis is that morphological variation carries some signal of topical variation for terms in text. The observation we make supports this hypothesis: the most frequent case forms of Finnish terms are distributed significantly differently for query terms in relevant and non-relevant texts.

This skewness in distribution has potential to be used in several different ways. In a standard adhoc retrieval setting it has some potential for weighting query and index terms differently depending on their case and rôle in text of for the informed implementation of a generative approach such as is porposed by Kettunen. Alternatively, in future more semantically demanding applications, this distributional difference can be utilised to winnow out segments of text where informative terms e.g. occupy topical foreground or background, or where the narrative focus shifts with respect to some content bearing terminology.

6. Acknowledgements

We would like to thank Adam Blomberg and Johan Carlberger both at Euroling AB and SiteSeeker for their kind assistance during our experiments.

7. References

Chris Buckley and Ellen M. Voorhees. 2004. Retrieval evaluation with incomplete information. In *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 25–32, New York, NY, USA. ACM Press.

Hercules Dalianis and Bart Jongejan. 2006. Handcrafted versus Machine-learned Inflectional Rules: The Euroling-SiteSeeker Stemmer and CST's Lemmatiser. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May.

Bart Jongejan and Dorte Haltrup. 2005. *The CST Lemmatiser*. Center for Sprogteknologi, University of Copenhagen.

Kimmo Kettunen and Eija Airio. 2006. Is a morphologically complex language really that complex in full-text retrieval? In *Advances in Natural Language Processing*, pages 411–422. Springer Verlag, Heidelberg.

Kimmo Kettunen, Eija Airio, and Kalervo Järvelin. 2007. Restricted inflectional form generation in managmenet of morphological keyword variation. *Information Retrieval*, 10:415–444.

Kimmo Kettunen. 2007. *Reductive and Generative Approaches to Morphological Variation of Keywords in Monolingual Information Retrieval*. Ph.D. thesis, University of Tampere, Department of Information Studies, Tampere, Finland.

Kimmo Koskenniemi. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. Ph.D. thesis, University of Helsinki, Department of General Linguistics, Helsinki, Finland.

Carol Peters, Martin Braschler, Julio Gonzalo, and Michael Kluck. 2003. *Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002. Rome, Italy, September 19-20, 2002. Revised Papers*. Lecture Notes in Computer Science , Vol. 2785. Springer Verlag.

Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck. 2004. *Comparative Evaluation of Multilingual Information Access Systems 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, Revised Selected Papers*. Lecture Notes in Computer Science , Vol. 3237. Springer Verlag.

M F Porter. 1980. An algorithm for suffix stripping. *Program*, 14:130–137, July.

Pasi Tapanainen and Timo Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 64–71.

Precision (Finnish)						
	fi-mo	fi-mm	fi-oo	fi-om	fi-so	fi-sm
P5	0.2375	0.3447	0.2842	0.2857	0.3415	**0.4222**
P20	0.1250	0.2021	0.1342	0.1571	0.1768	**0.2111**
map	0.0697	0.2780	0.1002	0.1286	0.1957	**0.2908**
bpref	0.0858	0.2988	0.1217	0.1547	0.2289	**0.3145**
Recall (Finnish)						
Finnish	fi-mo	fi-mm	fi-oo	fi-om	fi-so	fi-sm
Average recall	0.0406	**0.3736**	0.0903	0.0426	0.2173	0.1320
Queries with relevant hits (of 75)	16	**47**	19	7	41	18
Number of retrieved documents	136	**1553**	277	181	964	619
Retrieved relevant documents	40	**368**	89	42	214	130
Precision (German)						
	de-mo	de-mm	de-oo	de-om	de-so	de-sm
P5	0.3099	**0.3450**	0.2381	0.2800	0.2879	0.3392
P20	0.1077	**0.1325**	0.0798	0.1160	0.0985	0.1323
map	0.1203	**0.1560**	0.1294	0.1303	0.1200	0.1475
bpref	0.1395	**0.1835**	0.1561	0.1642	0.1418	0.1738
Recall (German)						
	de-mo	de-mm	de-oo	de-om	de-so	de-sm
Average recall	0.0732	**0.1015**	0.03207	0.05553	0.06223	0.1000
Queries with relevant hits (out of 100)	71	**80**	42	50	66	79
Number of retrieved documents	422	**522**	261	347	379	**522**
Retrieved relevant documents	153	**212**	67	116	130	209

Table 2: Finnish and German evaluation results

	Singular Nominative Genitive Partitive	Singular Inessive Illative Elative	Plural Nominative Genitive Partitive	
	Observed distribution			Sum
Relevant texts	3713	**547**	895	5155
Non-relevant texts	13292	**2966**	3277	19535
Sum	17005	3513	4172	24690
	Expected distribution			Sum
Relevant texts	3550	**733**	871	5155
Non-relevant texts	13454	**2779**	3300	19535
Sum	17005	3513	4172	24690

Table 3: Search term case distribution in relevant and non-relevant texts (the most divergent values marked in bold; χ^2 : $70.155; df = 2; p < 0.005$)