Compound terms and their constituent elements in information retrieval

Jussi Karlgren
Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, Sweden
jussi@sics.se

Abstract

Compounds, especially in languages where compounds are formed by concatenation without intervening whitespace between elements, pose challenges to simple text retrieval algorithms. Search queries that include compounds may not retrieve texts where elements of those compounds occur in uncompounded form; search queries that lack compounds will not retrieve texts where the salient elements are buried inside compounds. This study explores the distributional characteristics of compounds and their constituent elements using Swedish, a compounding language, as a test case. The compounds studied are taken from experimental search topics given for CLEF, the Cross-Language Evaluation Forum and their distributions are related to relevance assessments made on the collection under study and evaluated in terms of divergence from expected random distribution over documents. The observations made have direct ramifications on e.g. query analysis and term weighting approaches in information retrieval system design.

1 What is a compound?

Compounding is one of the basic methods of word-formation in human language. Two or more base word elements which typically occur independently as single words are juxtaposed to form a more complex entity. The compound elements can be concatenated without space, joined with a hyphen, or form an open compound with white space in between: "classroom", "cross-lingual", "high school". Compounding is a productive process: new compounds can be formed on the fly for ad-hoc purposes to treat topical elements in the discourse at hand. The semantics of a compound is typically related to the constituent elements, and most often the former constituent modifies the latter. Compounding has been studied in detail although not always in terms of function by linguists, terminologists, grammarians, and lexicologists over the past years; there are excellent overviews available for most any language one might be interested in. Compounding processes may show great surface differences between languages. Some languages use script systems that make no discernible difference between compounds and happenstance or syntactically motivated juxtaposition – ideogram-based Asian scripts, such as Japanese or Chinese, e.g. Some languages show a preponderance of open compounds and are restrictive in forming new closed compounds, such as the English language (see Quirk et al. (1985) for a comprehensive treatment of English compounding). Other languages again, such as Swedish, a near relation of English both in terms of cultural and linguistic history, tend towards closed compounds – with no white space between elements (see Noréen (1906) for a comprehensive treatment of Swedish compounding).

Compounds that originally are formed on the fly are eventually lexicalized and gain status as terms in their own right in a language. Terms such as "staircase", "blackbird" or "doorjamb" are not dynamically constructed for the purpose of a single discourse session or a single text – they are single simple words from the perspective of the language user. Compounds can also be borrowed from and loaned to other

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 111–115
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

languages and then often lose their character as a composite term: "homicide" is only in some sense an English compound. Other types of derivation such as affixation also resembles compounding to the extent that it may be difficult to draw a line. Is "eatable" a compound of "eat" and "able"? These types of process make compound analysis a demanding task for language engineering applications. When is it motivated to segment a compound to make it understandable and when should it be understood to be a lexical item in its own right?

2 Compounds in information retrieval

In information retrieval the problem of matching compounds to its separate elements has sparked some recent interest. Since compounds are typically formed topically, the constituent elements are quite likely to have topical focus in the text at hand. In languages where compounding often results in closed compounds (a "compounding" language) this will pose a problem for typical string search based retrieval systems: the query terms may contain a crucial element buried inside a compound, or alternatively, the index entries for some document may lack crucial constituent elements if they occur only or mainly inside compounds. In the last few years of growing interest in cross-lingual and multi-lingual information retrieval, several recent research efforts have addressed aspects of compounding and decompounding for the purposes of information retrieval.

A retrieval system tailored to the requirements of a compounding language would thus ideally split compounds both when indexing and when processing query terms. This has been tested by Braschler and Ripplinger (2004), who performed a set of information retrieval experiments on German text collections using various approaches to morphological analysis including decompounding, and found, much as expected, that decompounding efforts improved retrieval results considerably.

For indexing Swedish material splitting compounds at indexing time could be expected to improve recall by allowing queries to find elements that otherwise would be hid inside compounds. In experiments, Ahlgren (2004) has found that splitting compounds and indexing documents for both the entire compound and each constituent element separately yielded no significant effects, but in his experiments he did not attempt to decompound the query terms. In other experiments, Cöster et al. (2004) found that judicious splitting and expansion of query terms provided promising, if not entirely convincing results. Both index and query are likely to need decompounding to deliver practically useful results.

In a cross-lingual context, languages may have different compounding strategies. Translating from Swedish – tending towards closed compounds – to English – tending towards open compounds allows the strategy of splitting compound query terms and then translated element by element to formulate an English query to retrieve documents from an English index. Since most compounds in English are open, this strategy does not require any separate treatment of the index: the separate constituent elements are mostly reasonably elements of the query in terms of what the index contains. Hedlund (2003) has tested this strategy with productive and successful application to structured query construction, where the structure of the compounds can be translated into a structure of disjunctions and conjunctions of single term elements. It is evident that the structure of the compound itself carries information — which should not be discarded out of hand but instead be utilized in the analysis.

3 Experiment methodology

This present study is not an information retrieval experiment in the traditional sense. Most of the cited research efforts above have performed large-scale experiments using a retrieval system of their choice or of their own construction, and report aggregated results from several retrieval runs. However, most also seem to feel the need to supplement their results with a more detailed performance analysis, analyzing term and constituent element occurrences using an implicitly set theoretic approach: "the term XY occurred only rarely in the relevant documents whereas the constituent element X was fairly frequent". For the purposes of this study, no retrieval system was employed at all and only the primary data, term (and constituent element, as the case may be) occurrences are reported.

Terms may occur in texts for various reasons. Some occurrences are incidental and thus spurious for the purposes of topical retrieval; other occurrences are focussed, topical, and salient for topical retrieval systems. In this present study an arguably drastic simplified approximation to topicality is employed: if a term occurs in a document *more than once* it is considered topical; if the term occurs only once it is considered incidental. This is based on a three-parameter model of term distributions defined by Katz (1996): it postulates among others the parameter γ – the probability a term is topical or not. Katz' γ is estimated by the observed relative frequency of it appearing at least twice:

$$\gamma = \frac{N - n_0 - n_1}{N - n_0}$$

where n_0 is the number of documents without the term, n_1 is the number of documents with the term exactly once, and N is the total number of documents under consideration.

Katz' γ estimate here constitutes an estimate of the bound of usefulness of terms for information access analysis. The assumption, following Katz, is that if a term tends to reoccur in a document, it will tend towards topical use. Topical use can then presumably be utilized by a search algorithm, if well crafted.

4 Data

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.

The document databases consisted of the CLEF collection of Swedish newsprint for the years 1994 and 1995, the sixty Swedish-language topics for the 2003 evaluation cycle (numbers 141-200; only the title and description fields were used for this study, as in most experiments performed on the data), and the relevance judgements given for those topics. The document collection consists of some 140 000 documents of news articles in Swedish, most rather brief.

5 Occurrence statistics for compounds

In Swedish running text between a tenth and a fifth of tokens are compounds: the material in the CLEF collection seems to tend towards the higher figure. Search queries can be assumed to be more topical than other types of textual material, and compounds can accordingly be expected to be more frequent in information need specifications. Statistics on CLEF topics bear out this assumption. Taking the sixty Swedish CLEF topics from 2003 we find about one thousand term token occurrences. Once morphological variants are conflated and stop words are taken out (including query specific terms such as "find" and "document"), we find that out of the somewhat less than four hundred individual terms used in the topics more than ninety are compounds, and that if duplicates are removed the numbers are even more striking: almost every second unique noun is a compound. (Compound analysis is always a question of judgments: some analyses are debatable, other compounds are missed. Hand checking by several assessors indicates these errors cancel each other out for the purposes of this study.) As a comparison, the English versions of these topics contain only two or three closed or hyphenated compounds. The statistics as shown in table 3 establishes beyond any doubt that for compounding languages such as Swedish queries should be understood to be compound dense. Any query analysis for a compounding language procedure should reflect this fact.

6 Goal task

The following statistical observations are meant to guide the task of query generalization. Given that a query contains a set of terms, some of which are compounds $A+B$, would the query be enhanced by adding further terms to it, e.g. constituent elements such as A or B?

7 Topical relevance and compounds

If compound query terms are frequent, they should be expected to be frequent in the target document set as well. Documents that are assessed as relevant with regard to the query topic in question contain compound terms in

Table 1: Occurrences of compounds in one year of Swedish CLEF topics.

	Tokens	Two-place compounds	Three-place compounds
All	392	78	14
Unique	333	77	14

general with about the same frequency that other documents in the collection, and the occurrence of single terms and compounds is not dramatically different between relevant retrieved and non-retrieved documents.

The constituent elements do not behave symmetrically, however. As can be seen in table 7 former elements taken from a compound query term more often to be topically used in relevant documents than latter elements or new compounds based on either former or latter elements. This is consistent with arguments in the philological literature (Quirk et al., 1985; Noréen, 1906) where the former element of a compound is observed to be a focal component in an implicit predication.

These statistics show that on average it would seem to be advisable to add the former element of the compound to the query: if the query contains term "diamantindustri" (*diamond+industry*) "diamant" is likely to be a useful query expansion, far more than other "industri" would, and more than new compound terms such as "diamantring" (*diamond ring*), "oljeindustri" (*oil industry*), or "bilindustri" (*automobile industry*).

8 Remaining questions to study

8.1 Distributional overlap

While the preceding set of statistics indicate that compounds can profitably be analyzed and elements treated individually, the overlap between compounds and their elements remains a factor in determining the informational value of constituent elements vs entire compound: if the overlap is large, the marginal gain from introducing new terms can be assumed to be of low utility.

9 Principled prediction of topical overlap

The above argument is based on the assumption that compounds follow a general pattern. The generalization may be useful and practical but is likely to obscure underlying systematic differences between different types of compounding processes and different types of constituent element. The question is whether it is possible to predict the likely utility of adding constituent elements to a compound query term by observing the relative distributional statistics of the constituent elements and the compound in the collection without involving relevance assessments or human intervention.

The statistical treatment of compound elements is a special case of the general question of terminological topical interdependence.

Questions that are being investigated in continuing research efforts aim at the prediction of topical characteristics based on observed distributional characteristics. They include questions such as: Can we make a principled choice of elements to generalize from? Are elements in certain positions more valuable than others? Can we use characteristics of the bare elements to make that choice? Can we use total overlap to predict usefulness of constituent elements?

10 Discussion

The results of the present study are unequivocal on one level of analysis. As previous studies have shown, compounding languages should be treated as such. This study confirms that observation.

Additionally, this study observes that compounds are not simple and happenstance juxtapositions of equally salient elements. Compounding is a mechanism used by authors and speakers for a reason, and this can be usefully utilized in the analysis of e.g. information retrieval queries. It would appear to be worth the trouble to select constituent elements by their distributional characteristics and by the structure of the compound term rather than by their appearance as an unanalyzed compound in a query.

Thirdly, on a methodological level, this study claims that the set theoretical and distributional methodology used, while yielding less immediate results in terms of ranked retrieved document sets, gives better purchase for imple-

Table 2: Katz' γ for compound term (AB) elements in the target collection.

	all documents	relevant documents	N
Single query terms	0.253	0.368	295
Compound query terms (A+B)	0.215	0.273	92
Former element alone (A)	0.276	0.383	94
Latter element alone (B)	0.270	0.338	99
Former element recombined (A*)	0.235	0.335	232
Latter element recombined (*B)	0.267	0.362	117

mentation in various systems rather than evaluation based on the quirks and characteristics of large scale information retrieval systems, however competent and useful the systems are for the task they are built for.

References

Per Ahlgren. 2004. *The Effects of Indexing Strategy-Query Term Combination on Retrieval Effectiveness in a Swedish Full Text Database*. Ph.D. thesis, Department of Library and Information Science, University College of Borås.

Martin Braschler and Bärbel Ripplinger. 2004. How effective is stemming and decompounding for German text retrieval? *Information Retrieval*, 7:291–306.

Rickard Cöster, Magnus Sahlgren, and Jussi Karlgren. 2004. Selective compound splitting of Swedish queries for boolean combinations of truncated terms. In Carol Peters, Martin Braschler, Julio Gonzalo, and Martin Kluck, editors, *Fourth Workshop of the Cross–Language Evaluation Forum (CLEF 2003)*. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany.

Turid Hedlund. 2003. *Dictionary-Based Cross-Language Information Retrieval: Principles, System Design and Evaluation*. Ph.D. thesis, Department of Information Science, University of Tampere.

Slava Katz. 1996. Distribution of content words and phrases in text and language modelling. *Natural Language Engineering*, 2:15–60.

Adolf Noréen. 1906. *Vårt språk*, volume 7. CWK Gleerup, Lund, Sweden.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. Longman, London, England.