Performance of Two Statistical Indexing Methods, with and without Compound-word Analysis

Introduction

In Germanic languages, compound words are very common and very productive. There are compound words which are bound and lexicalized and loose their semantic content when split (e.g. *albatross* or *jordgubbe*). This category will be referred to as *opaque compounds*. The opposite of the opaque compounds are the *productive compounds*, whose parts keep their semantic value when separated (Bjarnadóttir 2003). Among these are the compounds that are used, and sometimes invented, for a special context (e.g. *indexeringsmetod*).

Orthographically split compounds in Swedish are considered ill formed from a normative point of view. But within the domain of information retrieval the productive compounds give the individual word frequency lesser value than if they were two separate words. Therefore, it would be interesting to explore how statistical indexing methods perform when productive compounds are split as compared to when they are not.

In the study, two different indexing algorithms were analyzed. In order to evaluate the algorithms, operating with and without the addition of a split compound module, their performances were compared to the manual indexing of 30 students of linguistics at Stockholm University. All together, 15 news articles of different length and subject, from the Stockholm-Umeå Corpus (Ejerhed et al. 1992), were indexed manually and with two statistical indexing methods, both with and without the splitting of productive compounds. Indeed, a considerable deal of the manually indexed terms turned out to be compound words. As for the statistical indexing methods, the most successful one was equally successful with and without the split compound module, whereas the less successful method benefited considerably from the splitting of the compounds.

Background – indexing and natural language

Information Retrieval

Information retrieval (IR) embraces representation, storage, organisation, and access to information items (Salton and McGill 1983). IR has none whatsoever restriction on the format. But typically, retrieval systems include letters, documents of all sorts, newspaper articles, books, research articles etc. Sometimes IR only refers to the technical auxiliary tools such as database, index program or search (or matching) program, and not how to retrieve information (Sundström 1981).

Usually information retrieval is viewed upon as a circular procedure, where the user makes a request for information to a system and recursively evaluates the response until the information need is fulfilled (Berghem 1982). The system could be any kind of system, for example a card catalogue system in a library.

The user's request is compared with a description of stored items in the system. When the comparison is executed, the request is matched with the description of the stored data. Each

match means some item in the stored data corresponds to some item in the request. For example, a user wants to know something about jaguars (Manning and Schütze 2002). The user writes *jaguar* in a question box, with the hope of finding information about the feline jaguar, and submits the inquiry to the system. The system will then compare the word *jaguar* to see if the word matches any of the stored items chosen to describe a document. These items, or *index terms*, are stored in a special file, an index file. Any time the system finds the index term *jaguar* for some document it retrieves the document and presents it to the user. Usually the system makes some automatic relevance assessment of the documents retrieved. This can be achieved by some algorithm, for example the *Probability Ranking Principle* (Manning and Schütze 2002). Using PRP the documents will be presented to the user in descending order of estimated relevance.

But in the *jaguar* case the user will get documents containing information about the feline jaguar as well as documents which contain information about the automobile brand *Jaguar*. These kinds of systems do not perform the disambiguation task. Systems that index natural language inherit the ambiguities of natural language.

The matching procedure is an orthographical one. Therefore, when dealing with Swedish one should also consider the elaborate noun inflections (i.e. *indexeringsalgoritm(er|en|ens |ers|erna|ernas)*). Consequently, it is important to use a stemmer to generate stem forms. Furthermore, Swedish is particularly inclined to produce new words through compounding (i.e. through *productive compounds*).

Statistical indexing methods

The aim of statistical indexing is to capture content words which have a good discriminating ability and a good characterizing ability for the content of a document (Sparck Jones and Robertson 1997). Discrimination ability means that the words are able to distinguish documents from one another. To capture the content of a document one talks of word characterization ability (Salton and MGill 1983).

Before the actual indexing takes place, a few normalizing processes apart from the tokenizing have to be performed. Stop lists are often used to strip the document from function words as prepositions, conjunctions etc. Another process has to do with the identification of a word's stem. There are to different ways of solving this problem – one is to use a stemmer and the other one is to identify the lemma forms of the words. The difference between these two techniques is that a lemma-identifier captures the real stem and the stemmer just guesses the word (Dura 1998). In SUC, the lemma form of every word is provided and we can easily replace the form in the text with the SUC lemma form (Ejerhed et al. 1992).

Most of the automatic indexing methods start with observing word frequency in natural language. In addition, one can observe a words frequency in a balanced corpus (e.g. SUC). It has been established that the distributional pattern of word types in natural language is irregular (Zipf 1949; Schultz 1968).

Words or terms, which occur in few documents, are considered more valuable to the content in a document than terms that occur frequently in several documents (Salton and McGill 1983). The terms that occur in few documents are regarded as being more informative of a text's content (Manning and Schütze 2002).

Inverse Document Frequency, or *Collection Frequency Weight*, uses this phenomenon to extract words that could describe a text's content. *IDF* is one of the statistical indexing methods that are used in the present study. In the study, the IDF formula is combined with the *Term Frequency* (*TF*) formula. TF is used to compute the frequency for each word in a specific document (Manning and Schütze 2002). This frequency value is supposed to catch how salient a word is for the document. When using TF it is important to use a length normalizer, otherwise the length of a document will affect the IDF value (Moens 2000).

The *Document Frequency* (*DF*), another important value, is necessary to know before computing the IDF. DF is used to compute, for each word type in a document corpus, how many documents of the corpus that contains the word. If a word occurs in few documents, it is said to be a good discriminator.

IDF could be computed in different ways. In the study, the following formula is used:

$$tf_i * \text{Log}(N/df_i) / \sqrt{\sum (tf_j * \log(N/df_j))^2}$$

(Moens 2000:94)

tf_i, the frequency for one word type **$_i$** in a document
df_i, the total sum of documents where word type **$_i$** occurs
N, the total sum of documents in the corpus
tf_j, the frequency for each word type **$_j$** in a document
df_j, total sum of documents where each word type **$_j$** occurs

To identify the word types in each document that should be singled out as index terms, one uses a threshold value based on test data (Viestam 2001). IDF should be recomputed every time a new document is brought in to the corpus.

An alternative to IDF is a model that is based on distribution, *Term Distribution Model*. This model has other ways to determine whether a word has the ability to describe the content of a document (Manning and Schütze 2002). In the study, a *Rank-frequency Distribution Model* (called *Luhn model* here) was used. The model applies *Zipfs law*, which says that the ability of a word to characterize a text is proportional to the words frequency in the text (Zipf 1949). This is described as the "Principle of Least Effort" – the writer or speaker uses the least vocabulary possible to express her-/himself. The Luhn model operates on Zipfs law by using words frequency value from the reference corpus and multiplies it with the words ranking value in the corpus.

$$\text{Log(Collection Frequency)} * \text{ranking} = \text{constant}$$

(Moens 2000:90)

Luhn established (Schultz 1968 (Luhn 1958)) that not only the most frequent words, like *the* or *of*, are bad index terms, but also words with very low frequency. Therefore, Luhn suggested that the deciles containing the lowest and the highest frequency words of a document, be cut off the indexing procedure. This means that 80 percent of the words in a document will be held as index terms.

In this study, dynamical thresholds are used for both indexing methods. The method based on Luhns assumptions will be used according to his suggestion. The indexing method based on

IDF (in combination with TF) will select from an article those words that correspond to the 80 percent highest values assigned to each one of the word types in the IDF computation.

Morphological compound structure

The statistical indexing methods operate on words frequency in different ways. But neither of them takes special account of specific language phenomena like productive compounds or specific homographs. Swedish, for instance, has the noun homograph *dom*, meaning 1: *cathedral*, 2: *verdict* or *judgment*. Both these meanings could be good candidates for index terms. Productive compounding is a very common phenomenon in Swedish. But before we will be able to distinguish between the different definitions and classes of compounds we have to unfold some morphological terms.

Swedish morphological units can be subdivided into free morphemes and bound morphemes, which in its turn can be divided into derivative morphemes, inflective morphemes and joint morphemes (*fogemorfem*) (Malmgren 1994). Free morphemes are independent words with independent meaning, also called root morphemes. Free morphemes usually belong to the open word classes such as noun, adjective, verbs. The opposite of free morphemes are the bound morphemes, also called grammatical morphemes. Bound morphemes are not entire words, one could say that they modify the free morphemes and give them a more or less different meaning. Inflective morphemes, joint morphemes and derivative morphemes are bound morphemes. The joint morphemes constitute a class of grammatical morphemes that is important for compounding in Swedish. This class is the glue between two free morphemes. There are five graphemes representing the joint morphemes – a, o, u, s, e – but the joint morpheme is far from always applied.

barn|s|lig (*childish*)	(root morpheme + joint morpheme + derivative morpheme)
av|led|ning (*derivative* (n))	(derivative morpheme + root morpheme + derivative morpheme)
barn|s|lig|are (*more childish*)	(root morpheme + joint morpheme + derivative morpheme + inflective morpheme)
av|led|ning|en (*the derivative*)	derivative morpheme + root morpheme + derivative morpheme + inflective morpheme)

English morphological theory identifies compound structures in English even if the words are not joined together, for example *window cleaner* and *emergency sail change*, the motive for this being phonological (Spencer 2001). Swedish compounds also demonstrate phonological features. The Swedish equivalent to Spencers term *compound stress* is *sammansättningsbetoning* (Riad 1997).

Two English definitions of what a compound word is:

Compound words are new words formed out of other words, e.g. black bird, girlfriend, babysit, supermarket parking lot attendant, emergency sail change

(Johnson 2002)

A compound noun consists of two or more words used together as a single noun. The parts of a compound noun may be written as one word, as separate words, or as a hyphenated word.

(Holt, Reinhart and Winston 2003)

One Swedish definition of what a compound word is:
A compound word is a word which can be split into at least two word-like units, which both of them contain at least one root morpheme [author's translation] (Malmgren 1994:32)

Although the phonological criterion goes for Swedish compounds as well, the search for a normatively split compound is a search in vane. Compound words, in Swedish, are generally viewed as an orthographic joint unit between to root morphemes.

The most common compounds in Swedish are combinations of noun plus noun, adjective plus noun, and verb plus noun (with descending frequency). The combination verb plus verb is very rare (Malmgren 1994). Within the class of noun compounds there are combinations with proper names and other encyclopaedic units, for example *mellanösternspecialist* (*Middle East specialist*), *Hultsfreds-biljetter* (*ticket to Hultsfred* (music festival)) and *Björnborgväska* (*a bag of the brand Björn Borg*). This kind of compound is quite common in Swedish (Järborg 1998).

The meaning of a compound cannot always be predicted by its parts. In such cases one has to simply know the meaning of the compound to understand the word. This type of compound is called *opaque* or *exocentric compounds*. The Swedish word *jordgubbe* (*strawberry*) is a typical opaque compound and *blackboard* (which is not just any black board) is an English example. Another class of compounds is the class of *productive compounds*. These are usually compounds that a writer creates for a specific context, e.g. *indexeringsmetod* (*indexing method*) (Ekeklint 2001). But there are also compounds in this class which are high frequency words used frequently in every day spoken language, e.g. *lastbil* (*truck*). For the present study, we are mainly interested in those compounds that are created for a special context.

Problems with splitting compounds
When splitting Swedish compounds one has to know where the parts start and stop (Dura 1998). It can be quite difficult to split compounds at the right place. Sometimes the joint morpheme and the duplication letters coincide, for example *glassko* could mean *glass shoe* (*glas | sko*), *ice cream cow* (*glass | ko*) and *ice cream shoe* (*glass | sko*). Swedish spelling conventions do not allow more than two identical letters following each other. Word-final gemination of a letter will be reduced to a single letter if, when compounding, the geminate sequence meets with a word starting with the same letter.

Sometimes bound morphemes coincide with homographic free morphemes. The compound *självständighetsförklaring* (*declaration of independence*), for instance, could (erroneously) be split into five free morphemes in two different ways: *själv|ständig|het|(s)|för|klar|(ing)* or *själv|ständing|hets|för|klar|(ing)*, when in fact at least *het*, *för*, *ing* and the joint morpheme should be analyzed as bound morphemes.

Method

Fifteen news articles, from SUC, were indexed both manually and automatically. The manual indexing was done by 30 students of linguistics at Stockholm University. Each news article was indexed by to different students. The students were requested to choose 10 content words. These words should not be proper names, geographic names or company names. If the students thought this was very important they were permitted to choose 5 extra words.

When the articles were indexed automatically, they were indexed twice with each of the two indexing methods – that is both invoking and not invoking the split compound module. The main rule for splitting the compounds was formulated thus: if anyone of the parts of a compound was a noun, adjective or a verb, the compound was split, otherwise it remained a single orthographic unit. This rule was subsequently implemented in the automatic split compound module. For the design of the programs, the reader is referred to Andersson (2003).

Evaluation frame

In order to evaluate if the split compound module did give a positive effect on the automatically indexing procedure, the results statistical methods were first compared to each other accounting for the compound splitting variable. Thereafter the statistical methods were compared with the manual indexers, who served as reference for ideal indexing.

The data were confronted with the following questions:

1. How many index terms were selected by the statistical indexing methods as compared to how many terms that were selected by the human indexers?

2. How many index terms are index terms for more than one article?

3. How many terms have the human indexers chosen as important for the content of an article, that the statistical methods have left out, or have not found?

4. How many of the compound words that were chosen by the human indexers have also been chosen by the statistical indexing methods?

Results

The students indexed the articles with an average of 18.7 terms per article. An average correlation of 3.5 terms was observed comparing the pair wise manual indexing of one and the same article. That is, merely a sixth part of all manually selected index terms for an article was common to the two students indexing the same article.

In the manual indexing, an average of 6.3 index terms per article was compound words. This means that almost a third of the index terms were compound words.

The statistical indexing methods, with and without the split compound module, are referred to as *Luhn* and *IDF* respectively. Considering the salient variable of the study – the splitting of the compounds – the statistical indexing will yield four different outputs: *Luhn, Split Luhn, IDF* and *Split IDF*.

1. How many index terms were selected by the statistical indexing methods as compared to how many terms that were selected by the human indexers?

All the automatic indexing methods generated more than ten times as many index terms as the human indexers. Split IDF chose the greatest number of index terms (in average 279

words per article). Luhn (without split compound module) demonstrated the lowest number of terms – an average of 228 words for each article.

2. How many index terms are index terms for more than one article?

This frame question tries to show how well the statistical indexing methods perform in distinguishing between different articles. IDF indexed an average of 118 terms per article that also occurred in at least one other article. The corresponding values for Split IDF, Luhn and Split Luhn were 144, 117 and 134 respectively. Luhn shows the best result from this point of view. Nevertheless, this means that at least 70 of the terms chosen for an article have been chosen for some other article as well.

3. How many terms have the human indexers chosen as important for the content of an article, that the statistical methods have left out, or have not found?

This evaluation frame question measures how well the statistical indexing methods have performed in capturing the **relevant** index terms. The performance of the statistical indexing methods was compared to the manual indexing. IDF and Split IDF managed to capture 96 percent of the relevant index terms, whereas Split Luhn and Luhn only captured 86 percent and 68 percent respectively. Thus both IDF and Split IDF demonstrate the best characterizing capacity. On the other hand, Split IDF also selects the greatest number of unnecessary terms.

4. How many of the compound words that were chosen by the human indexers have also been chosen by the statistical indexing methods?

IDF and Split IDF both managed to capture almost 99 percent of all the compound words indexed by the student reference group, which is an average of 6 words for each article. Split Luhn performed second best, capturing 95 percent of the compound words, whereas Luhn only indexed 57 percent of the compound words chosen by the reference group.

Discussion

Although Luhn has fewer index terms for each article and fewer words that are index terms for more than one article, it performs considerably poorer in capturing the relevant words. In conclusion, merely because a statistical indexing method performs well for some criteria such as selecting relatively few terms or not producing overlapping index terms, this does not mean the method is a good one.

For an automatic indexer it is important to capture the word that really matters for the searcher, i.e. the words that best describe the content of a document. Luhn would have needed a higher threshold value to perform better in capturing the relevance word. On the other hand, IDF probably would have benefited from a lower threshold value, which would have cut off some of the irrelevant terms.

For IDF and Split IDF the split compound module did not increase the performance of the method, each indexing method capturing 96 percent of the relevant terms. As for the Luhn methods, the split compound module increases the relevant term rate from 68 percent to 86 percent, which is an excellent improvement. As we have seen, both IDF and Split IDF

managed to capture 99 percent of the relevant compound words. To verify that these results are more generally valid, that is that the effect of a split compound module is dependent on what kind of indexing method is used, one has to examine a bigger corpus and also optimize the threshold values for each indexing method.

The manual indexing reveals that almost 34 percent of the chosen index terms were compounds. Compounding makes expressions concise and shorter than a phrase:

kameldrivare = person who herds camels

midsommardans = dance associated with the midsummer night festivities

bilbarnstol = special security chair for children, used in automobiles = (*car*) *baby seat*

bilbarnstolsbälte = (*car*) *baby seat belt*

Therefore, even if a split compound module does not improve the performance of an index method, it could be of interest to split compounds for the users ease. When searching for information on some topic, it is not always easy to guess which contextual compound the writer has come up with. Contextual compounds do hide good index terms, for example *missil* (*missile*) in *missilvapen* (*missile weapon*), *nyhet* (*news*) in *nyhetsrapprtering* (*news report*) and *artist* in *svensktoppsartist* (an artist who is associated with a special music list on Swedish radio).

But it is not always good to split compound words – some opaque compounds look like productive compounds, for example *jordgubbe* (*strawberry*), its parts being *jord* (*earth*) and *gubbe* (*old man*). For some high frequency productive compounds, it is questionable if they should be split, for example *lastbil* (*truck*) and *trappsteg* (*step, one distinctive part of a staircase*).

One way to identify the relevant contextual compounds is to look for independent occurrences of its parts in the text, particularly the second part. This approach is based on the assumption that the writer uses the contextual compounds so frequently that she/he chooses to omit parts of a compound and only refer to the concept represented by the compound via one of its parts, usually the right most part of the compound. A text where the writer always uses the complete contextual compound could be tiring for the reader. In fact, it is like reading a text where the writer does not use pronouns for nouns and proper names.

References

Berghem, Agneta (1982) *Datorbaserad informationssökning vid Foa4* C 401157-C1, A3, B1

Bjarnadóttir, Kristin (2003) *Searching for Compounds: The Representativeness of Corpora* GSLT Graduate School of Language Technology

Dura, Elzbieta (1998) *Parsing Words* Doctoral Thesis, Göteborg University.

Ekeklint, Susanne (2001) *Tagga samman - ett verktyg gör semantisk analys av svenska sammansättningar,* Master's Thesis in Computational Linguistics, Göteborg University Available online at:
http://www.cling.gu.se/theses/finished.html
(2003-10-24)

Ejerhed, Eva & Källgren, Gunnel & Wennstedt, Ola & Åström, Magnus (1992) *The Linguistic Annotation System of the Stockholm-Umeå Corpus Project – Description and Guidelines* Department of Philosophy and Linguistics, Umeå University

Holt, Rinehart and Winston (2003) *Elements of Language – 3rd course* pp 375-378 Available online at:
http://www.hrw.com/language/eolang/peonline/course_3/ch12/lg1312375_378.pdf
(2003-10-15)

Johnson, Mark (2002) Handouts for class 2002, *CG41 Morphology, the structure of words.* Brown University. Available online at:
http://www.cog.brown.edu/~mj/classes/cg41/handouts/wk02a.pdf
(2003-10-15)

Järborg, Jerker (1998) *Sammansättningssemantik* Rapport 1 från LBAB (Lexikal betydelse och användningsbetydelse) Department of Swedish, Göteborg University

Malmgren, Sven-Göran (1994*) Svensk lexikologi - ord, ordbildning, ordböcker och orddatabaser* Lund, Studentlitteratur

Manning, Christopher D and Schütze, Hinrich (2002) *Foundations of statistical natural language processning.* Massachusetts Institute of Technology, 1st edition 1999, 2nd edition with corrections 2000.

Moens, M. (2000*) Automatic indexing and abstracting of document texts USA*. Kluwer Academic Publishers

Riad, Tomas (1997) *Svensk fonologikompendium* Department of Scandinavian Languages, Stockholm University

Salton, Gerard and McGill, Michael J (1983) *Introduction to modern information retrieval* USA, McGraw-Hill Inc.

Schultz, Claire K. (1968) *H. P Luhn: pioneer of information science selected works* London, American Documentation Institute

Sparck Jones, K and Robertson S. E. (1997) *Simple, proven approaches to text retrieval* Department of Information Science, City University & Computer Laboratory. University of Cambridge.

Spencer, Andrew (2001) "Do English have productive compunding?" In: *Preceedings from 3rd Mediterranean Morphology Meeting* Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Barcelona September 2001

Sundström, Erik. (1981) *Detta är datorbaserad informationssökning* Lund, Studentlitteratur

Viestam Susanne (2001) *Three methods for keyword ex traction.* Master's thesis, Language Engineering Programme, Department of Linguistics, Uppsala University
Available online at:
http://stp.ling.uu.se/~matsd/thesis/
(2003-10-24)

Zipf, George Kingsley (1949) *Human Behavior and the Principle of Least Effort*. Massachusetts, Addison – Wesley Press Inc. Cambridge