The Automatic Indexing System AUTINDEX

Øivin Andersen and Peter Widell
University of Bergen and Politikens Forlag, Copenhagen

In this presentation we will describe how a set of basic programmes for information retrieval can be used to generate lists of key words and short reference abstracts from authentic documents in machine readable form. The programmes used are basically along the line of Salton and McGill's book **Introduction to Modern Information Retrieval** from 1984.

The theoretical background which lead us to problems connected with information retrieval was a test of van Dijks semantic reduction rules, the so-called "macro-rules" as described in his book **Macrostructures** from 1980.

From a linguistic point of view van Dijk's theory can be seen as a theory of a part of human language competence in Dell Hymes extended sense. Important aspects of this competence is man's ability to delete, to generalize over and to construct compressed information on the basis of a series of connected facts. A basic question is: How can this ability be explicated? According to van Dijk this can be achieved by a set of semantic reduction rules. These rules do not operate on the surface structure of a given text, but on a deeper, propositional structure, which van Dijk refers to as a **micro text base**.

The problems involved in converting a text from its authentic form to a propositional form are numerous, and have been extensively discussed in philosophy and modern semantics over the years. We will not discuss them here, although they are far from being solved yet.

In modern textlinguistic analysis it is common to view the underlying meaning structure of a text as a hierarchy of content units. Van Dijk's micro text base is seen as a hierarchy of propositions where the communicatively more "essential" propositions occupy a higher position in the hierarchy, dominating the less essential propositions.

This base is the data on which van Dijk's macrorules operate. The task of the macrorules is to define the global theme of the text by the operations DELETION, GENERALIZATION and CONSTRUCTION. DELETION deletes the propositions which are not relevant to the interpretation of other propositions in the micro text base. GENERALIZATION abstracts from the semantic details by constructing more general propositions, using a hierarchically structured lexicon. By CONSTRUCTION the more condensed proposition is established by using what van Dijk somewhat vaguely refers to as a **cognitive set**, which is the summary writers' world knowledge making up the interpretative background of the text. A series of propositions are substituted by a more general one, based on prototypicallity and stereotypes. The details of these rules with examples can be studied in van Dijk's book.

All these rules are operations which are reductive in the sense that the denotative content of a text is reduced to a condensed form. Moreover, the operations are cyclic, which means that a given text will have several variants with different degrees of reduction.

Van Dijk's theory of macrostructures was tested by us and evaluated both in a linguistic and an information retrieval context.

In the context of information and documentation there are two basic types of information reduction applied on LSP texts: Firstly, short summaries used for quick reference of the basic, global topic of a text (so called "abstracting"), and secondly, a short index list containing the most essential key words (so called "automatic indexing").

From a linguistic point of view the question is whether van Dijk's theory is an adequate model of man's ability to make abstracts from coherent texts. The linguistically based theory should contain a method for generating summaries which are acceptable to a competent abstract writer. But the linguistic aspect of this is closely connected to the information retrieval aspect of the question. As Richmond points out it is reasonable to expect that, instead of quantitative data on how many people found what they wanted and by what means, it would be more to the point to consider how well models of information retrieval systems have matched the way in which humans process information (Richmond 1982:153). In the ideal world both these aspects should coincide. In other words, efficiency in information retrieval should give us a clue to the usefulness of a textlinguistic theory as a model of man's information retrieval faculty, like the efficiency of a machine translation system should give us a hint as to its usefulness as a linguistic theory serving as a model of man's ability of translating. In that case our theory should contain a method which enables us to mirror various user groups' level of knowledge about the topic of the given text, and their motivation for seeking information in the text.

Our test of van Dijk's theory revealed that the lack of rule ordering was problematic, since manipulation with ordering would give us non-coherent abstracts. But the most important weakness of the theory was that the **conditions** for the applications of the rules were not given. Especially the concept **cognitive set** had to be given an empirical content. It turned out that the reduction rules were partly based on intuition which needed explication and testing. Moreover, the "translation" from authentic text to algorithmic form is problematic, specially the convertion of the conditions of described situations from implicit to explicit form, i.e. the set of implicit semantic relations and tacit background knowledge of the world as fundamental aspects of the construction of a micro text base of a given text.

Nevertheless, these problems do not exclude the possibility of constructing programmes that satisfy our demands to key words and abstracts in a relatively simple way. The notorious problem of representing tacit background knowledge has to some extent been taken care of in AUTINDEX by the construction of fragments of a thesaurus which combines prototypes and stereotypes with indexed words from texts.

In the following we will describe the AUTINDEX programme. The programme design was developed by us, and Widell has implemented it in VEDIT PLUS.

Automatic indexing has been known and applied for some years and various systems have been developed. There is a good survey in Salton et al 1984. The basic guidelines of construction are constant, and AUTINDEX are constructed along the same basic ideas. The main purpose of indexing is to give rapid and reliable access to relevant texts in a text corpus.

The performance of automatic indexing systems is usually measured in terms of **recall** and **precision**. To put it simply, "recall" is the amount of retrieved relevant material, and "precision" is the lack of noise in the retrieved relevant material.

One important type of automatic indexing is called free text searching. Various efficiency tests have been carried out on these systems. They revealed that low precision and recall is often the result of an insufficient mumber of search words. Moreover, these systems are unable to express important semantic relations between central key words in a text.

Problems concerning polysemy and synonymy cannot be handled in a simple manner, and topically important information which is implicit in texts cannot be retrieved.

Finally, there is a problem related to LSP texts. A salient property of LSP texts is the high degree of specificity. Finding adequate search words for such texts is a very difficult task for the user. As Blume (1987:32) points out texts must be combined with a structured thesaurus, since the user cannot be expected to be familiar with the terminology of the subject area represented by the texts.

A solution to these problems is the use of an interactive thesaurus which indicates the semantic relations between search words, using scope notes to disambiguate the words. Among other things it can perform document weighting to increase precision.

Two basic procedures are always present in automatic indexing systems:

Firstly, an initial filtration where the text runs through a series of reduction filters (stop lists) where high frequency words, also called function words, are deleted

Secondly, a structuring of the text material by using a battery of statistic computations giving a word stem list with indicated frequencies.

A basic condition for manipulating texts in this manner is the availability of an electronic dictionary. AUTINDEX is based on such a dictionary: **Nudansk Ordbok**, a contemporary dictionary of Danish from 1990.

An important aspect of AUTINDEX is its use of a thesaurus, which to a considerable degree increases the quality of the selected index words of the investigated text. In AUTINDEX it is possible to supplement an index list which is generated by statistic computation with new relevant index words taken from the thesaurus.

But AUTINDEX not only uses a thesaurus. It also allows updatings in the applied thesaurus. Index words from the investigated text may be suggested as new entries in the thesaurus. Consequently, AUTINDEX can be said to contain a **documentary based, interactive thesaurus**. Since this facility in AUTINDEX is a general facility, not dependent on any specific thesaurus, we can say that AUTINDEX to a considerable degree facilitates the construction and updating of thesauruses in general.

Texts can be measured as to their ability to discriminate text type or genre. If an indexing term for instance has a high score in documents relating to a particular topic, but a low overall score in the total collection, the term is said to have a high **discriminating value**.