EMBEDDED SUBLANGUAGES AND NATURAL LANGUAGE PROCESSING

Richard Kittredge

Université de Montréal

Abstract

Most recent systems for the large-scale intelligent processing of natural language texts are designed to accept only a restricted variety of language. In certain cases this restricted subseto of the language constitutes a sublanguage, for which it may be possible to write a relatively precise and compact sublanguage grammar. Several research groups are currently exploiting the restrictions in scientific and technical sublanguage grammars for tasks such as information retrieval and automatic translation.

At present the notion of sublanguage is rather loosely defined, or variously understood. There is no rigorous procedure for determining whether an arbitrary corpus of texts belongs to the same sublanguage, or deciding what the limits of grammaticality and acceptability for a given sublanguage are. The tendency has been to assume that texts which are produced in similar communicative situations and which refer to some delimited universe of discourse must be from the same sublanguage. This assumption is reinforced by the discovery of special structural properties, lexical collocations, etc. in the sentences and texts of a functionally and semantically homogeneous corpus. Not only is the vocabulary of a specialized sublanguage more limited than the vocabulary of the general language, but the very special lexical co-occurrence restrictions are equivalent to a statement of the possible properties and relations between the objects of the universe which is the subject matter of the sublanguage in question.

A detailed study of several sublanguages of English and French has focussed on certain aspects of text structure and lexical selection which are important for automatic processing of sublanguage texts as well as for a clarification of the defining properties of sublanguage. . It is known that a sublanguage text may have different subsections which differ sharply in their sentence structure, textual linking devices, etc. An obvious case in the division in technical manuals into descriptive and procedural (maintenance) subsections. More elusive is the sub-division of certain scientific texts into (1) a science language component, describing the properties of objects and relations between them in the domain of scientific inquiry and (2) a meta-science component, describing the relationship of the observer to his methods and results. Such a division has been pointed out (but not studied in detail) both by Harris (1968) in his theoretical discussion of sublanguage and by Sager (1972) in her study of the sublanguage of experimental pharmacology. The two components may be present in the same sentence, as when a clause of science sublanguage is embedded under a meta-science predicate. Such embeddings also show up in weather synopses. Their proper analysis is a prerequisite to the choice of proper parsing strategy and dictionary construction for the automatic analysis of these texts.

A different kind of subdivision can be found in the language of stock market reports where a sublanguage describing the trading activity on various stock exchanges can be embedded in a much broader (sub)language describing the world of government decisions, international events and the like, and their effects on the market. The way in which this embedding occurs is related to the level of user of the text. Typically, however, causal sentences are adjoined to the right of narrow sublanguage sentences under special conjunctions in a way suggestive of a right ideal in a ring, as outlined in Harris' theoretical study. The narrow sublanguage portion of the text is distinct, both lexically and grammatically, from the right-adjoined sentences from the broader language set. A parser must be able to recognize these junctures and exploit the restrictions found in the narrow (embedded) sublanguage without failing to parse the sentences of the loose matrix.

The junctures between embedded sublanguages and matrix portions of text can also be viewed in terms of the linking properties found in both subsections. The scope and type of linking found in each subcomponent may be quite distinct. Rules of anaphora must be constrained to operate within a particular subcomponent.

The proposed paper will present and discuss several of the above-mentioned properties

observed in the embedded sublanguages of stock-market reports, meteorological synopses and mathematics texts. Some conclusions will be drawn for the relationship between various types of sublanguages and the language as a whole, and implications discussed for the design of intelligent text-processors.