3. Concept Extraction

Christine Montgomery, Chairperson

Logicon, Inc.
Woodland Hills, CA 91365

Panelists

Robert Moore, SRI International
Robert Simmons, University of Texas
Norman Sondheimer, Sperry Univac
Robert Wilensky, University of California at Berkeley
Wiliam A. Woods, Bolt, Beranek and Newman

3.1 Introduction

The panel on concept extraction from natural language text defined a variety of different views on the problem at issue. First, we considered what would be the properties of a very powerful natural language understanding system which would serve as an ultimate model for our deliberations, but was clearly not achievable within the five-year time frame specified as the primary area of interest to the Navy for the purposes of this workshop. However, it could serve as a goal toward which progress could be made in a series of approximations, beginning within the next five years and extending beyond.

The "ultimate" system we envisioned was a model for understanding the text of scientific and technical literature in order to discover technological innovations, discern evolution of technological trends, and predict where science and technology will evolve in the future. Such a system would consist of two major components:

1. Natural language understanding expert(s), whose task would be to read and assimilate the narrative text of scientific and technical literature.
2. Science and technology analysis expert(s), whose function would be to evaluate the content of individual articles as well as the cumulative contents of the collection of scientific and technical materials, and to identify trends and innovations, using models of current developments in science and technology that define what would constitute an innovation, a trend, etc., and that specify a variety of other rather inexplicit types of knowledge used by human experts to evaluate and predict.

With this lofty goal somewhat vaguely identified as a Holy Grail for the future, we concentrated our efforts on defining approximations to that goal, selecting as our focus the military message analysis and routing problem presented by NRL personnel as representative of a particular set of problems the Navy is currently attempting to solve. Within this context, the panel addressed two types of development issues:

1. How to manage the evolution of a type of system based on existing message routing technology to a more capable, knowledge-based system within the next five years.
2. How to manage the evolution of the next generation knowledge- based system for message routing from the last development stage reached in (1).

These two stages of development are discussed in Sections 3.2 and 3.3, respectively.

3.2 The Next Five Years

In planning the evolutionary development of existing military message routing systems into more capable, knowledge-based systems within the next five years, the panel considered a variety of techniques and approaches that might be added to existing systems to augment the current level of capability and "grow" the system in the direction of a more powerful natural language processing model.

There are two experimental development methods which could be used to achieve this goal. The first would be to begin with an actual operational message routing system, install some version of this system at NRL as an experimental vehicle[2] and use it as a basis to build on, adding the components which provide the incremental capabilities discussed below, and systematically evaluating the effect of the additional capabilities. An alternate approach would be to construct a first stage experimental message routing system without using any existing system as a foundation and proceeding as indicated above.

[1] For example, the predecessor company of Logicon's Operating System Division – Operating Systems, Inc. (OSI) – built a variety of such systems for several different computer configurations. Since these systems were constructed under government contract, they are available at cost to the government.

3.2.1 First Stage: Add State-of-the-Art CL Components

The initial stage of improved capability can be achieved by incorporating into the basic message routing system described above several state-of-the art computational linguistic components which are almost characterizable as "off-the-shelf". Specifically, these include a spelling corrector, a morphological analyzer, and a synonym generator. The spelling corrector would be of a type similar to the capability currently available under the UNIX operating system, tailored for the Navy message routing application (that is, including a variety of Naval acronyms, relevant proper nouns such as place names and ship names, and other special jargon terms). Intelligent morphological analyzers have been built for a number of natural language processing systems; these can be supplemented by current non-numeric processing techniques for partial match of character strings.

A synonym generator, on the other hand, while well within the state of the art, still requires development of the knowledge base of synonyms in some manner. To illustrate, suppose that the system must automatically produce for a user all the natural language synonyms for the terms in a routing profile or retrieval query of a user who wants all messages about "military aircraft in the vicinity of the Red Sea". In order to generate the synonyms for each of those terms (which users should then be allowed to prune or supplement as they prefer), the system must either have been primed beforehand by an extensive development of a thesaurus-type knowledge base, or the system must have an ability to acquire such knowledge from users in an incremental manner (where the burden of developing the knowledge base of synonyms is mainly on the user), or a combination of both (for an illustration of the latter, see Katter and Montgomery 1972).[2]

3.2.2 Second Stage: Augment Key Word Based Concept Extraction

The second stage of improved capability in the basic message routing system would be achieved by augmenting the current key word base content analysis component in several steps. It should be noted that these steps are based on the in-depth analysis of messages and the routing environment described in Section 3.3, which should thus be conducted in parallel with the development of the first stage improvements discussed in the preceding section.

[2] The operational message routing systems described in footnote 1 allow users to develop synonym tables and utilize those produced by others; however, these are not automatically triggered when a routing profile is entered, but must be specifically requested by the user.

Content analysis of messages based on key words involves a rudimentary identification of concepts, with no attempt to identify relations between concepts. Thus a message reporting on shipments of cotter pins from Cuba to Albania would not be distinguishable from a similar one reporting on shipments of cotter pins from Albania to Cuba. Consequently, the initial step in augmenting the ability of a message routing system to analyze the content of messages and extract concepts from text would be the introduction of grammar rules and a syntactic analysis component. A semantic grammar approach in which the categories of the grammar map fairly directly into user information requirements ments would constitute a productive initial capability. Such a grammar could be built along the lines of the LIFER parser which was incorporated into LADDER (Burton 1976) specifically to answer questions about ships at sea.

A second development step in managing the evolution of a key word based concept extraction system to a natural language understanding system involves the construction of taxonomic hierachies which provide the system with the knowledge to attempt rudimentary inference, allow property inheritance, and exercise some very simple analogic reasoning based on the taxonomies. The work of Woods on constructing and exploiting taxonomic hierarchies is illustrative of this type of capability. The methodology used by SRI's NANOKLAUS system for organizing knowledge has a similar basis (Haas and Hendrix 1980).

A further enhacement would involve the utilization of more general inference about the content of the message (for example, rule based inference, where the concept of rule includes frames and scripts as well as production rules). Many messages relate information about some particular type of event that has occured many times before and will continue to occur, e.g., a missile or space launch. Knowledge about the entities and relations involved in such an event can therefore be codified in terms of rules or scripts which explicate the expected contents of a message. This provides the basis on which the system can derive information which is not explicity stated in the message and can identify elements of information which are unexpected or lacking. The experimental message understanding system described in Silva et al. 1979 is an example of this type of capability, as is the work of Simmons and Chester (to apppear, 1982).

The final developmental level deemed to be achievable within the next five years differs from the levels previously discussed in that it incorporates one type of meta-knowledge, that is, knowledge that is not concerned with the content of the messages, but with the goals of the recipient of the message. For example, the recipient may have gotten the message simply for information; on the other hand, the recipient may be

required to take some action on receiving the message rather than merely filing it away, and the actions he may take could be quite different depending on the content of the message (e.g., information that an activity has begun, or that an activity which has been going on for some time is continuing, or has ceased). Thus the consequences of the natural language understanding failing to identify the type of content and causing the message to be misrouted can be quite serious in an operational environment.

3.2.3 Third Stage: Provide for Coping with the Limitations of an Augmented Operational/ Experimental System

Before the augmented message routing system could be used even experimentally by persons other than its designers, it must have the means for dealing with its own inadequancies. Suppose the lexical lookup encounters a word which is not in the dictionary, or the parser cannot analyze an input construction, even though all the words are known. Approaches to solving the problem of unexpected inputs fall into two categories, and a robust system will incorporate some partial solutions of both types. The first type of approach involves what heuristics and other reasoning strategies the system can attempt without appeal to the user; the second involves user interaction.

What the system can do without appealing to the user in the case of an unknown word ranges from:

1. very simple heuristics which assume a misspelling and attempt to identify the word by partial match strategies, to
2. the complex strategy of utilizing a protype knowledge structure to deduce that the unknown word must be an entity of type x because it fits an otherwise unsatisfied slot in such a structure that requires an entity of type x.

Some productive approaches to handling structures that are ungrammatical from the system's perspective involve systematically relaxing some of the grammatical constraints built in to the parser (e.g., subject/verb agreement), as discussed in Weischedel and Black 1980.

In appealing to the user for assistance, a critical factor is supplying not only the right information about the particular failure, but the right amount of information presented in words that the user can understand, and with suggested remedial actions that are general from the system perspective, not confusing to the user, and unlikely to elicit erroneous inputs. Accomplishing all this is a non-trivial problem.

3.3 The Next Generation

Assuming that all these thorny problems of the message routing system to be developed in the next five years are solved, the issue of managing the evolution of the last stage of that system into something like the powerful natural language understanding system described in Section 1 remains to be considered.

An important part of this evolution, which was alluded to above, is the specification of the two major components of such a system, namely

A. A Natural Language Understander, which will be based on a model of the universe of discourse represented in the content of the messages;
B. A Message Routing Expert, which will be based on a model of Navy organizational and functional structure.

In the following discussion, Section 3.3.1 describes the research tasks for component (A), while Section 3.3.2 treats the tasks for component (B).

3.3.1 Research Tasks for the Natural Language Understander

In order to provide for the evolution of the system described in the preceding section to a next generation, knowledge based message routing system – as well as to furnish basic information for the development to be achieved in the next five years – the initial research task should begin immediately, and the second research task should begin as early thereafter as feasible. The initial five year development effort does not depend on the remaining research tasks, which are aimed at the next generation system. The research tasks are described in the sections which follow.

Message Analysis

The initial task consists of an in-depth analysis of the domain of Navy messages that the routing system must handle. The objective of the task is to provide detailed data on the content and format of the messages, that is, information about the lexical items, jargon terms, classes of lexical items representing concepts, the concept hierarchy, the types of relations which can exist among concepts, the types of syntactic constructs that those relations map into, and the discourse structures of different types of messages.

Tractability Analysis

This research task will define three classes of messages, in terms of their tractability to automated processing, i.e., messages which are

1. Tractable for the initial stage of development (Section 3.2.1);
2. Tractable within the next five years (that is, tractable at least in terms of the stage of system development described in Section 3.2.3);
3. Tractable in the future.

Selection of Knowledge Representation

Since this is one of the most debated aspects of building knowledge based systems, it seems unnecessary to discuss it further here. Any novices to this issue will find Ron Brachman's SIGART 1980 survey one of the more interesting recent attempts to document the thinking of various research groups on the subject. Moore's thoughtful article on a theory of logical forms for natural language expressions is refreshing in that it focuses on the content rather than the form of knowledge representation (Moore 1981).

Development of Knowledge Structures for Message Subset

This task involves the development of knowledge structures for a selected subset of the universe of discourse represented by the messages. The task thus incorporates several subtasks, all of which can be considered as recommended research areas to be pursued.

Lexicon

The development of a lexicon for the message sublanguage is a complex task because of the many acronyms and jargon terms used; some of these duplicate common English words, but with quite different meanings, (e.g. "sinker", as in "sub went sinker") and others must be analyzed in order to furnish appropriate input to the parser and other system components (e.g., "center spa 226k hawk 9 mi"). See the report of the Panel on Sublanguages for further details.

Conceptual Structures

This task involves the construction of appropriate higher level conceptual structures (e.g., propositions, frames, scripts) which embody the intensional and extensional knowledge of the system.

Inferential Component

The inferential component uses the system's knowledge base to derive implicit information, to generalize and revise beliefs, and to attempt to understand the intentions of the actors whose actions are reported in the messages. The last area is an extremely complex one in which little research has been done. Wilensky's work on the goal-oriented behavior of actors in stories is a notable contribution (Wilensky 1978).

Discourse Analysis Component

Most of the relevant issues for the subtask involving definition of this component are covered in the preceding papers in this article. To recapitulate these issues:

- tense/time
- anaphoric reference
- focus tracking (updating of short term memory)
- goal analysis (analysis of goals of actors in the message, goals of the originator of the message, and goals of the recipient of the message)
- local and global inference, identifying discourse coherence
- significance assessment (identifying the most important and/or most salient concepts in the discourse, from the various goal-oriented perspectives mentioned above)

Integration of Knowledge Sources into a Coherent Model

Finally, all the various sources of knowledge alluded to above must be integrated into a coherent model of the message domain, which can be manipulated by analytical procedures operating autonomously and in parallel, but under integrated control.

3.3.2 Research Tasks for the Message Routing Expert

Simultaneously with the initial research task for the Natural Language Understander, the first task described below should be initiated.

Analysis of Selected Naval Command

An organizational and functional analysis of a Naval command selected as initial users for the near term (five year development system) must be performed in order to understand in general terms the goals of message originators and recipients.

Formalization of Expert Knowledge Involved

This task involves the detailed specification of the goals of message originators and recipients.

Relation of Content of Messages to Goals of Message Users

The final task calls for the mapping of message contents to user goals and the definition of a mechanism to accomplish this routing function, which is the goal of the systems described in Sections 3.2 and 3.3.