Automatic Indexing and Government-Binding Theory

Robert J. Kuhns
205 Walnut Street
Brookline, MA 02146
USA

ABSTRACT

This project note describes a system that receives, parses, indexes, and routes news reports. The core of this automatic indexer is a parser based on Government-Binding Theory which derives thematic and binding relationships of arguments of the sentences of stories. These syntactic structures are interpreted by a semantic processor which is linked to conceptual representations of terms from a controlled indexing vocabulary. As a result, the system is capable of indexing news with respect to a large set of terms that denote the content of the articles.

BACKGROUND

With the rapidly increasing volume of text being generated, transmitted, processed, and stored, it becomes critical that information retrieval and routing be highly efficient, both in time of processing and accuracy. To this end, indexing techniques have become the primary focus of much research, and yet these methods have relied on automatic keyword identification from texts. This is not to say that natural language techniques have not been examined with respect to their relevance for indexing and retrieval (cf. /Sparck Jones and Kay 1973/, /Walker, Karlgren, and Kay 1977/, and more recently, /Salton and Smith 1989/). It is that most systems rely on the presence or absence of keywords with additional mechanisms such as proximity constraints, statistical weighting, word-stem truncation, and boolean retrieval expressions. However, these methods do not take into account the syntactic and semantic structure inherent in the text being indexed. That is, they make virtually no use of the fact that it is natural language and not a collection of arbitrary strings of characters that is being processed.

Natural language processing (NLP) can make its most valuable contributions to those aspects of indexing where the keyword approaches fail, viz., the assignment of terms to text based on their semantic or conceptual content. This involves deriving abstract relationships among conceptual units. For example, consider a story stating:

(1) China bought 6,000 tonnes of wheat from the United States.

One plausible categorization of (1) is that it is about foreign trade. However, the phrase "foreign trade" does not appear in (1), and it is invariably absent from foreign trade stories in general. Furthermore, it is extremely unlikely that such foreign trade stories could be retrieved in an efficient manner, i.e., with a few simple queries. The central issue is that although the particulars (e.g., country names and types of commodities) vary, the basic meanings of foreign trade stories are equivalent at some level, and that this level is valuable for indexing purposes. This suggests that systems that could operate at a conceptual level would be capable of indexing in ways that could permit highly effective retrieval.

It is with the assumption that NLP technology can provide the means of categorizing text that guide several recent efforts. In particular, each of /Hayes et al. 1988/, /Kuhns 1988/, and /Rau and Jacobs 1988/ describes systems that characterize news reports with results that could not be obtained by keyword methods alone. Since a news analysis system (NAS) was first reported in /Kuhns 1988/, a number of major enhancements to its design and underlying functionality have been incorporated. It is the purpose of this paper to report on the current state of NAS and its components.

SCOPE AND OBJECTIVES

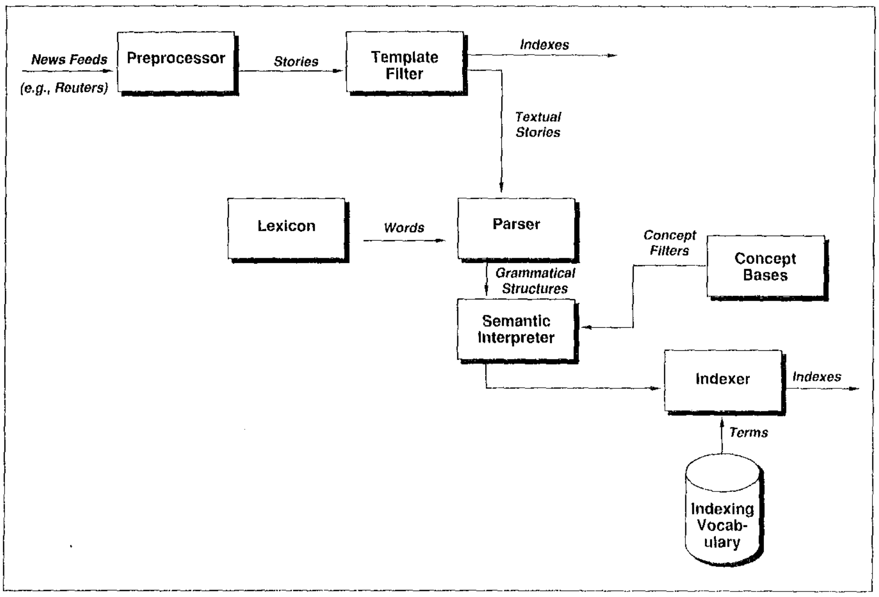
A primary design goal behind NAS is to develop a system employing NLP technology that would be capable of either routing news from electronic news feeds in real-time or indexing news with respect to very large sets of indexing terms (authority files).[1] This vocabulary is broad and ranges over diverse domains, with terms representing proper names, concrete objects, or abstract relationships. The last type is dependent on the content of the stories and is particularly suited for the syntactic/semantic techniques of NAS.

The form of the indexes that NAS produces is a set of pairs of headings and subheadings or descriptors. The headings are frequently proper names while the descriptors add detailed information to the index by denoting various relationships of entities mentioned in the news reports.

ARCHITECTURAL OVERVIEW

The architecture of NAS is modular consisting of several main subsystems, viz., a set of preprocessors and template filters, a parser, a lexicon, a semantic interpreter, a set of concept bases, and an indexer (Figure 1). The system is transportable in that it can be interfaced to different news streams and indexing vocabularies.

A preprocessor receives a news stream which can be from a satellite dish link, a direct line, or a text file, and identifies the beginning and ending of stories in addition to their titles, sentences, and words. Since the format of each news feed, e.g., Reuters or Kyodo, is distinct from the others, a single preprocessor will accept only one news feed. For rigidly-formatted articles that are numerical or non-textual in form, a template filter, which is an indexing component of low-level routines, categorizes them from the title while deriving specifics from the body of the story.



An Architectural Overview of NAS
Figure 1

Figure 2 is an actual example (from Reuters) that has been indexed by the template filter. The system outputs the company name with a descriptor "3rd Quarter Earnings," as well as the current and cumulative earnings or losses.

@R101903647
/&ACTMEDIA INC <ACTM.0> 3RD QTR LOSS
WESTHAMPTON BEACH, N.Y., Oct 19, Reuter -
Shr loss 14 cts vs profit 10 cts
Net loss 1,674,000 vs profit 1,207,000
Revs 26.7 mln vs 19.1mln
Nine Months
Shr loss 19 cts vs profit 34 cts
Net loss 2,280,000 vs profit 4,080,000
Revs 71.6 mln vs 59.8 mln
Reuter

Indexes:

Company Name -	ACTMEDIA INC
Descriptor -	3RD Quarter Earnings
Subject -	Net loss 1,674,000
Descriptor -	Current Earnings
Subject -	Net loss 2,280,000
Descriptor -	Cumulative Earnings

A Numerical Story and Its Index
Figure 2

In contrast, textual stories require grammatical processing and these are sent to the parser and semantic interpreter. The parser which relies on the principles of Government-Binding (GB) Theory (/Chomsky 1981/), outputs predicate-argument structure of each sentence of a story.[2] In doing so, the parser identifies empty categories, viz., PROs, traces, and variables, and thematic relations, and resolves antecedent and anaphor and pronominal bindings. It should be noted that the parser is interfaced to a lexicon of over 17,000 items that was developed by analyzing strings (words) from a newswire. The size of the lexicon is sufficient for news processing.

The semantic interpreter maps the grammatical structures onto conceptual representations or filters stored in a concept base. For instance, a representation for "Product Introduction" is

(2) Agent (COMPANY)
Predicate (INTRODUCE)
Theme (PRODUCT).

This structure denotes that a product introduction is one where a company introduces (or, synonymously, releases) a product. In short, it is a list of typed nodes. A report will be characterized as a product introduction story if it contains a sentence some of whose grammatical components (e.g., agent, predicate, theme) can be associated to the corresponding nodes of (2). Suppose, for example, a news item reports

(3) Alpha Corp said it plans to release a new workstation in Japan.

The parser, in accordance with GB principles, produces:

(4) Agent (Alpha Corp)
Predicate (said)
Proposition
    Agent (Alpha Corp)
    Predicate (plans)
    Proposition
        Agent (Alpha Corp)
        Predicate (release)
        Theme (a new workstation)
        In (Japan)

The parser binds the pronominal it, the agent of plans, to Alpha Corp, the subject or agent of the matrix clause. The parser also detects an empty category, viz., PRO, in the embedded sentence (proposition) with release as the verb and binds the pronominal to it.

Since bound arguments share the same semantic features, the semantic interpreter determines that the agent of release in (4) is of type COMPANY. In other words, PRO inherits the property of COMPANY from the agent of the matrix sentence via the intermediate pronominal it. It also determines that the predicate release is synonymous with introduce and the theme workstation is a product. With the arguments typed and membership of the predicate within a synonym class known, the semantic processor can match the corresponding nodes of the most deeply embedded clause of (4) with (2), and thus determines that the sentence is about a product introduction. Associated with each conceptual filter is a set of indexing procedures that are invoked by the indexing mechanism when a conceptual filter is satisfied. These functions are integrated with databases containing the indexing vocabulary and they identify specific information about a story including company, personal, and product names, and descriptors indicating specific relationships. Figure 3 illustrates the corporate and personal name identification capabilities and the level of "understanding" as reflected by the subheadings.

@R080100252
/&ALCO HEALTH<AAHS.0> CHIEF EXECUTIVE RETIRES
VALLEY FORGE, Pa., Aug 1, Reuter - Alco Health Services Corp said Ray B. Mundt has been named acting chairman and chief executive officer, succeeding John H. Kennedy, who is retiring.

Indexes:

Company Name -	Alco Health Services Corp
Descriptor -	Officials and Employees
Personal Name -	Kennedy, John H.
Descriptor -	Retirement
Personal Name -	Mundt, Ray B.
Descriptor -	Selection\Appointment

A Textual Story and Its Index
Figure 3

BENCHMARKS

Formal benchmarks have been established based on news from Reuters. On a Symbolics 3640, NAS can process entire days of news (500-600 stories/day) in 35-40 minutes and can assign indexes to approximately 75% of the stories. (The planned goal of at least 85% coverage is certainly achievable.) Accuracy was judged extremely high by a group of independent indexers and editors. Quality could not be judged quantitatively due to the complexity and subjectivity of the indexing terms and procedures.

FUTURE DIRECTIONS

In addition to continual extensions to the various components of the system, a design for an interface of NAS to deductive databases has begun. The development of this extension would enable databases to be generated automatically with indexes being stored as logical relations, thereby, permitting retrieval or alerting capabilities based on explicit as well as implicit or inferred information.

A NOTE ON THE IMPLEMENTATION

NAS was developed in ZetaLisp on Symbolics workstations. It has been converted to Common Lisp and runs on MacIvory and Macintosh computers.

CONCLUSION

The results of NAS demonstrate that it is possible to employ natural language processing and, in particular, linguistic-based parsers, to extract conceptual information from texts (news). This coupling of theoretical (GB) results with a large-scale application provides insights into the possibilities and limitations of computational as well as formal linguistics.

NOTES

[1] The natural language component, however, is not restricted to news processing. With another preprocessor and links to an expert system, the system has processed text found on insurance applications.

[2] The use of a GB-based parser within this application results from the ongoing research and development on this type of parser. (/Kuhns 1986/ describes an earlier implementation of GB Theory.) From a research perspective, a parser based on linguistic theory and applied to "real-world" text helps identify the boundaries or interface conditions between core and peripheral aspects of the theory. In other words, since GB is a model of core grammar and language contains marginal or marked constructions, a GB-based parser must co-routine principles of the theory with language-specific rules in order to have wide coverage. (/Tomita 1988/ makes a similar observation.) Thus, it is this combination of a psychologically-real theory and an application using free text that may provide insight into the human sentence processing mechanism.

REFERENCES

Chomsky, N., (1981), Lectures on Government and Binding, Foris Publications, Dordrecht, Holland.

Hayes, P.J., L.E. Knecht, and M.J Cellio, (1988), "A News Categorization System," in Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas.

Kuhns, R.J., (1986), "A PROLOG Implementation of Government-Binding Theory," in Proceedings of the XIth International Conference on Computational Linguistics, Bonn, West Germany.

Kuhns, R.J., (1988), "A News Analysis System," in Proceedings of the XIIth International Conference on Computational Linguistics, Budapest, Hungary.

Rau, L.F., and P.S. Jacobs, (1988), "Integrating Top-Down and Bottom-Up Strategies in a Text Processing System," in Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas.

Salton, G., and M. Smith, (1989),"On the Application of Syntactic Methodologies in Automatic Text Analysis," in Proceedings of the Twelve Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, Cambridge, Massachusetts.

Sparck Jones, K., and M. Kay (1973), Linguistics and Information Science, Academic Press, New York.

Tomita, M., (1988), "Combining Lexicon-Driven Parsing and Phrase-Structure-Based Parsing," in Proceedings of the XIIth International Conference on Computational Linguistics, Budapest, Hungary.

Walker, D.E., H. Karlgren, and M. Kay (eds.), (1977), Natural Language in Information Science, FID Publication 551, Skriptor, Stockholm.