A Knowledge-Based Approach to Conceptual Information Retrieval from Full Text

Inien Syu
University of Central Florida

Sheau-Dong Lang
University of Central Florida

The conventional methods of full text processing, such as pattern matching and keyword indexing, may result in waste of rich information. Recently, the knowledge-based approach has been used in natural language text processing. Unlike past generations of information retrieval systems which dealt exclusively with abstracts, title headings, and keywords, today's intelligent information retrieval systems manipulate full texts such as technical manuals, reports, letters, and news articles. This paper discusses the design and implementation of a knowledge-based information retrieval system, concentrating on the issues of knowledge acquisition, knowledge representation, knowledge base refinement, and the information retrieval mechanism.

1. Introduction

The field of information retrieval has seen a growing interest in the natural language research community. Past generations of information retrieval systems, which have used the techniques of pattern matching, adjacency retrieval operators, and keyword indexing, are facing a major barrier to extract useful information from full texts (Blair, 198). Efficient retrieval from full text requires basic understanding of the text as well as indexing it. To overcome this barrier, the knowledge-based system approach is adopted. Many systems have been developed successfully in different domains, such as KRITON (Diederich, 1987), START (Katz, 1988), TRPQ (Syu, 1989), SCISOR (Rau, 1989, MedIndEx (Humphrey, 1989), and KBNL (Barnett, 1990).

There are three issues regarding the design of a knowledge-based system: knowledge acquisition, knowledge representation, and knowledge base refinement [Cohen, 1985). The knowledge-based system proposed in this paper adopts a domain-specific knowledge acquisition tool FBI (Frame-Based Interface) which helps a domain expert build a classification hierarchy as the initial knowledge base. A direct benefit of this approach is that the knowledge base is highly modular and modifiable. Following the trend of the knowledge-based approach for information retrieval systems in determining the main theme without necessarily paying attention to individual words (Mauldin, 1987), our knowledge-based system constructs noun phrases as the index terms for useful conceptual information. The knowledge base of our system also contains a set of rules for applying the domain-knowledge to retrieving index terms from full texts.

The remainder of this paper is organized as follows. Section 2 presents the knowledge acquisition tool FBI and the acquisition method. The knowledge base refinement is also described in that section. Section 3 presents a frame-based hierarchical representation which stores and accesses the semantic information of the natural language texts.

This manuscript was originally presented at the 1991 Information Resource Management Association International Conference.

Frame: thing
Subcategory-of:

Frame: physical thing
Subcategory-of: thing

Frame: idea
Subcategory-of: thing

Frame: inanimate
Subcategory-of: physical thing

Frame: animate
Subcategory-of: physical thing

Frame: animal animate
Subcategory-of:

Frame: plant
Subcategory-of: animate

Figure 1: Example of Evidence Provided by a Domain Expert

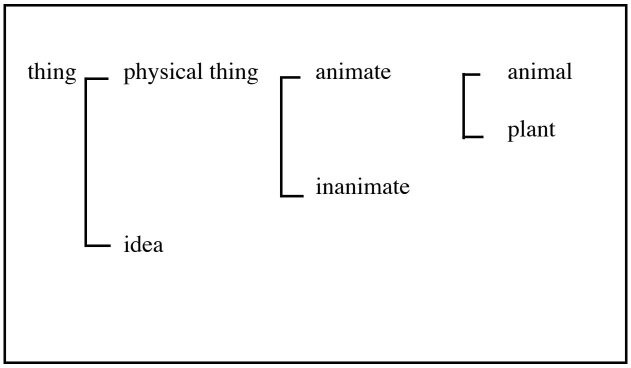

Figure 2: Hierarchy of Taxonomy Generated by FBI

Section 4 presents the mechanism of retrieving index terms. Section 5 concludes the paper and points out some future work.

2. Knowledge Acquisition

Knowledge acquisition is largely a matter of mapping the knowledge supported by humans into a representation recognized by computer systems (Kahn, 1987). In our system, a Frame-Base Interface (FBI) is designed for domain experts to initialize and refine the knowledge base. The output of FBI is represented in a frame-based hierarchy. The employment and use of the required knowledge is a major task for knowledge-based systems. Section 2.1 explains the transformation between the information of the domain experts and the internal knowledge representation of our knowledge-based system using FBI. Section 2.2 describes the utilization of the acquired knowledge. The knowledge base refinement is presented in section 2.3.

2.1 Acquiring Knowledge from Domain Experts

A Frame-Based Interface (FBI) is implemented to acquire the knowledge of a domain expert. FBI prompts the expert to list the evidence that are commonly relevant in the expert'sxomain. The expert is encouraged to be as specific as possible. The resulting knowledge base can be viewed as an under-specified hierarchy of nodes and links, where the nodes represent concepts and links represent relationships between concepts. To fully specify the hierarchy, a domain expert is requested to supply three kinds of information: (1) the category of each node, (2) the type of each link, and (3) the type and value of each attribute associated with a node.

FBI is developed using FRL (Frame Representation Language) to explore the use of frames as a knowledge representation technique. FRL allows properties to have defaults, comments, and constraints, and to inherit information from the same type (Roberts, 1977). Because of the nature of FRL and the information provided by the domain expert, FBI is able to classify the relationships among the nodes in the hierarchy. Figure 1 shows an example of evidence provided by a domain expert and Figure 2 shows the resulting hierarchy generated by FBI.

2.2 Utilization of Acquired Knowledge

Natural language analysis of noun phrase construction, which is chiefly needed in information retrieval, is an extremely difficult and challenging problem. Various attempts to come up with general rules for noun phrase understanding have been unsuccessful (Salton, 1990). However, in a specific domain with sufficient knowledge about the domain, this problem can be solved by the combination of syntax, semantics (domain knowledge), and production rules.

Our knowledge-based system uses a natural language parser to analyze and eliminate trivial syntactic ambiguities of the sentences. This parser is developed for incorporation into SEAN, a knowledge-based writer's aid system (Syu, 1990). An input text is processed by the parser using the ATN grammar rules and dictionary information. The output of this syntactic analysis is used by the semantics analyzer to generate terms suitable for indexing purposes.

It is clear that the syntactic analysis, which is not based on contextual and other considerations, can not be completely successful in eliminating syntactical ambiguities. To solve this problem, a semantics analyzer is designed to provide proper specifications for generating unambiguous index terms. This is accomplished by consulting the production rules predefined in the knowledge base in order to match the result of syntax analysis against the most specific nodes in the frame-based hierarchy generated by FBI. The production rules of our knowledge base and the details of both syntax and semantics analyses are described in section 4.

2.3 Refining Knowledge Base

The discovery of incomplete knowledge is one important feature of knowledge acquisition tools, e.g. MOLE (Eshelman, 1986) and KRITON (Diederich, 1987). Currently, FBI offers editing support for the user to correct improper information in the knowledge base, such as redundancy, incompleteness, and inconsistency. FBI also facilitates the growth and maintenance of our knowledge base. However, the domain experts need to refine the knowledge manually.

3. Knowledge Representation

Our knowledge-based system uses a frame-based knowledge representation. The use of frames for computerized structuring of knowledge is usually attributed to Minsky (1975). Frames are data structures that name a collection of related concepts. Each frame can be viewed as a node. A specific concept, named by the frame, may point to another concept. A frame concept is linked to another concept by a specific relation. A frame concept can inherit the attributes and their values from its ancestors according the hierarchical linkages.

The relations that link the frames hierarchically in our system are **subcategory-of** and **instance-of**. The following hierarchy of the **weapon** and the **M240 machinegun** concepts is an example of both **subcategory-of** and **instance-of** relations:

```
weapon
 |
gun
 |
machinegun
 |
M240 machinegun
```

These concepts are represented in the knowledge base using the following frames:

```
(FRAME: weapon
     (SUBCATEGORY-OF: null)
     (ATTRIBUTE: parts(.....), function(...))
     (DEFINITION: instrument for self-protection,....))
(FRAME: gun
     (SUBCATEGORY-OF: weapon)
     (ATTRIBUTE: parts(.....)))
(FRAME: machinegun
     (SUBCATEGORY-OF: gun)
```

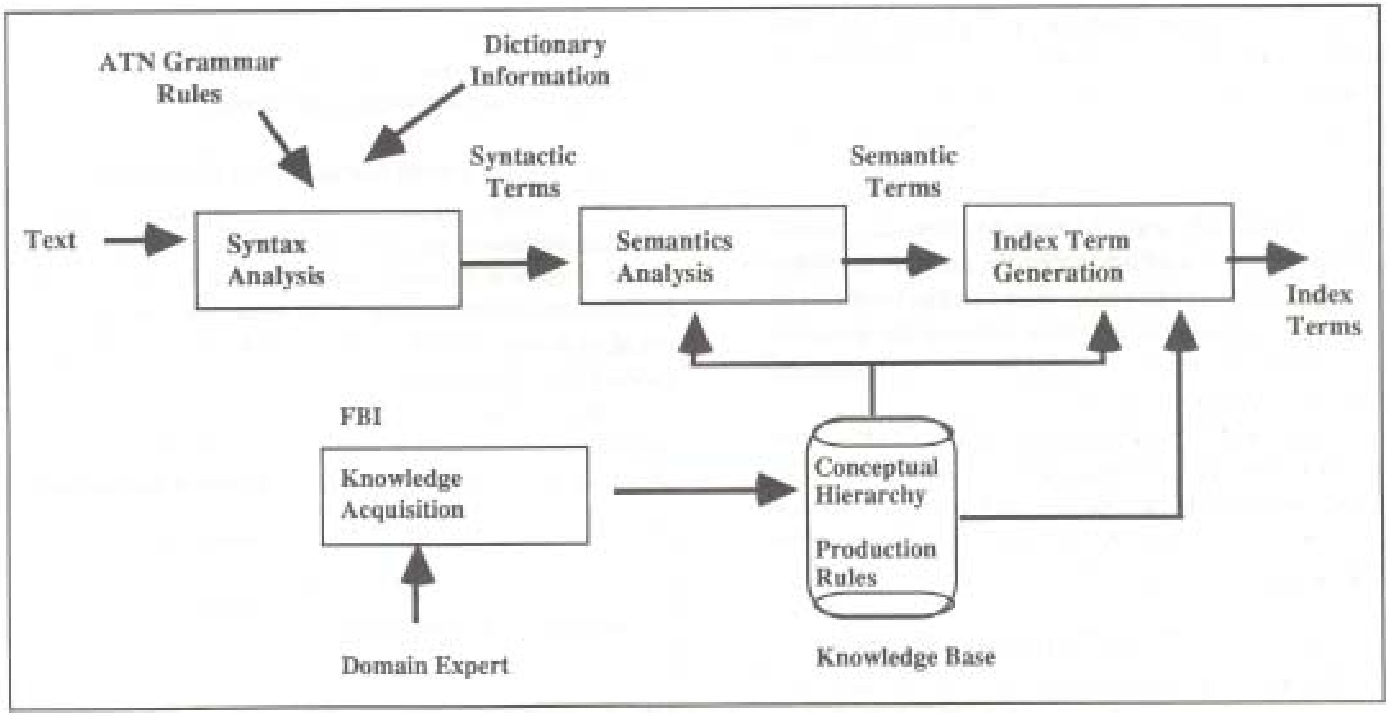

Figure 3: Task Flow of our Knowledge-Based System

```
Frame:weapon
            (Subcategory-of:  nil))
Frame: machinegun
            (Subcategory-of: weapon)
            (Attribute: user (gunner), operation (firing)))
Frame: machinegun1
            (Instance-of: machinegun)
            (Attribute: model (M60)))
(Frame: machinegun2
            (Instance-of: machinegun)
            (Attribute: model (M2)))
```

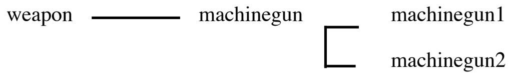

Figure 4: Frame-Based Hierarchy for the Sample Text

```
      (ATTRIBUTE: parts(.....)))
   (FRAME: M240 machinegun
      (INSTANCE-OF: machinegun))
```

The subcategory-of and instance-of relations, as in the preceding set of frames, provide the links whereby a frame accesses data from the ancestral frames. For example, the ATTRIBUTE and DEFINITION slots of the gun frame are inherited from the weapon frame. The frames are created and updated using FBI discussed in the previous section.

4. Using Knowledge Base for Retrieving Index Term

Most information retrieval systems use keyword-based indexing. In these systems, a list of keywords are provided and documents are analyzed for occurrences of these keywords. It has been proved that the presence or absence of keywords is not a sufficient criterion for analyzing documents, especially when they become too numerous (Maarek, 1989). The knowledge-based approach allows a deeper understanding of a text than the classical keyword-based approach. Our knowledge-based system is capable of extracting information of the relationships of the words involved. This information is necessary for distinguishing texts having similar keywords.

In section 4.1, we describe a set of production rules used for extracting conceptual terms from full text. The resulted conceptual terms characterize the text and stand for a description of the analyzed text. Section 4.2 presents the complete task flow of analyzing a text.

4.1 Index Term Extraction Rules

Our knowledge base contains a set of production rules, implemented in Prolog, that help extract index terms of a text. The left-hand side of the rules consists of one or more Prolog expressions which describe the properties of a retrieved term, and the right-hand side consists of a single Prolog expression that corresponds to the decision of selecting proper index terms. By examining the expressions on the left-hand side of the rules, a retrieved term is handled according to the decision expression of the right-hand side. The following show the production rules described in English. Our knowledge-based system runs these rules in a forward-chaining direction. An example of applying these rules is shown in section 4.2.

Marine, gunner, M60, and M2 machinegun team,
M60 machinegun, M2 machinegun, assistant gunner
machinegun, 1 minute, weapon, firing

Figure 5: Output of Syntax Analysis Stage

(Rule 1) If a term is the same as, or a synonym of, a concept (node) in the hierarchy, then accept this term as an index term.

(Rule 2) If a term is an attribute of a concept (node) in the hierarchy, then construct an index term by concatenating the name of the concept and this term.

(Rule 3) If an ancestor of a term is accepted previously as an index term, then construct a new index term by concatenating the ancestor term and this term.

4.2 Index Term Generation Procedure

Our system processes texts in three stages: syntax analysis, semantics analysis, and index term generation. Figure 3 depicts the task flow of our knowledge-based information retrieval system. To illustrate the tasks carried out in each stage, let us consider a sample text taken from a military education lesson.

The Marine acts as a gunner in a M60 and M2 machinegun team. Given a M60 machinegun, a M2 machinegun, and an assistant gunner, the Marine must mount the machin-

gunner, M60 machinegun, M2 machinegun,
machinegun, weapon, firing

Figure 6: Output of the Semantics Analysis Stage

```
machinegun gunner, machinegun M60, machinegun M2,
machinegun, weapon machinegun, weapon,
machinegun firing
```

Figure 7: Output of the Index Term Generation Stage

egun within 1 minute. The weapon must be ready for firing.

The corresponding domain knowledge represented in a frame-based hierarchy is shown in Figure 4. This sample text is used to illustrate the tasks carried out in each stage.

In the syntax analysis stage, an ATN grammar and a dictionary which contains complete words and their variations are used to output a list of noun phrases. These noun phrases, called syntactic terms, are retrieved without concerning the context and the word relationships. Figure 5 shows the output of this stage for the sample text.

In the semantics analysis stage, the syntactic terms which are not known in our knowledge base are rejected. The ambiguous syntactic terms are also modified to be as specific as possible. The knowledge contained in our knowledge base provides the basis for accomplishing these tasks. The resulting terms of this stage are called the semantic terms. Using the example in Figure 5, terms which are not in the knowledge base such as Marine, M60 and M2 machinegun team, assistant gunner, and 1 minute, are eliminated in this stage. Figure 6 shows the output of the semantics analysis stage.

In the index term generation stage, the semantic terms are checked by the production rules in the knowledge base. According to these rules, the semantic terms are concatenated with the names of the higher-level concepts in the hierarchy to specify their categories. The output of this stage are the final index terms extracted from the input text. Using the example in Figure 6, gunner, M60, M2 and firing are the attribute values of the machinegun frame. According to rule 2, index terms are generated by concatenating machinegun and each of these attribute values. In addition, the Machinegun frame is a subcategory of the weapon frame. According to rules 1 and 3, machinegun, weapon and weapon machinegun also are index terms. Figure 7 shows the final index terms retrieved from the previous sample text.

5. Conclusion

This paper describes a knowledge-based system for conceptual information retrieval from full texts. The key design of this system is a knowledge base of the semantics about a specific domain represented in the form of a frame-based hierarchy. A Frame-Based Interface (FBI) is developed to facilitate the acquisition of domain knowledge and refinement of the knowledge base. By incorporating the domain-knowledge with a set of rules predefined in the knowledge base, our system is capable of generating conceptual index terms.

This knowledge-based system has been adopted for the improvement of SEAN, a knowledge-based writer's aid (Syu, 1990) for authoring training materials written in controlled English. Using a domain-specific knowledge base, SEAN is capable of retrieving the theme of the paragraphs in a text. Future work includes applying to other natural language processing areas like automatic text classification, text summarization, and text translation.

Acknowledgement

The research reported in this paper is partially supported by a grant from the Florida High Technology and Industry Council (FHTIC), 1989.

References

Barnett, et al, (1990). Knowledge and Natural Language Processing, *Communications of the ACM, 33*(8): 50-71, Aug. 1990.

Blair, D.C. and Maron, M.E., (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system, *Communications of the ACM, 28*(3): 289-299, 1985.

Cohen, M.E. et al, (1985). Knowledge Representation and Classification of Chromatographic Data for Diagnostic Medical Decision Making, *Proceedings of the Second Conference on Artificial Intelligence Applications: the Engineering of Knowledge-Based Systems*, pp.481-486, 1985.

Diederich, J., (1987). Knowledge-Based Knowledge Elicitation, *Proceedings of the Tenth International Joint Conference on Artificial Intelligence,* Milan, Aug. 1987.

Eshelman, L. and McDermott, J., (1986). MOLE: A Knowledge Acquisition Tool That Uses Its Head, *Proceedings of the Fifth National Conference on Artificial Intelligence*, Vol. 2, 1986.

Humphrey, S.M., (1989). MedIndEx System: Medical Indexing Expert System, *Information Processing and Management,* Vol. 25, No. 1, pp. 73-88, 1989.

Kahn, G.S., (1987). From Application Shell To Knowledge Acquisition System, *Proceedings of the Tenth International Joint Conference on Artificial Intelligence,* Milan, Aug. 1987.

Katz, B., (1988). Using English for Indexing and Retrieving, Proceedings of RIAO 88', *Conference on Content-Based Text and Image Handling, MIT*, pp. 314-332, March, 1988.

Maarek, Y.S. and Smadja, F.A., (1989). Full Text Indexing Based on Lexical Relations, *Proceedings of the ACM SIGIR Conference*, pp. 198-206, 1989.

Mauldin, M., Carbonell, J. and Thomason, R.,(1987). Knowledge-Based Information Retrieval, *Proceedings of the 29th Annual Conference of the National Federation of Abstracting and Information Services,* Elsevier Press, 1987.

Minsky, M., (1975). A Framework for Representing Knowledge, In: Winston, P.H., editor. *The Psychology of Computer Vision,* New York: McGraw-Hill, pp. 211-277, 1975.

Rau, L.F., et al,(1989). Information Extraction And Text Summarization Using Linguistic Knowledge Acquisition, *Information Processing Management, 25*(4): 419-428, 1989.

Roberts, R.B. and Goldstein, I.P., (1977). *The FRL Primer, Massachusetts Institute of Technology Artificial Intelligence Laboratory*, July 1977.

Salton, G., Buckley, C., and Smith, M, (1990). On The Application of Syntactic Methodologies in Automatic Text Analysis, *Information Processing and Management*, Vol. 26, No. 1, pp. 73-92, 1990.

Syu, I. and Driscoll, J.R.,(1989). *A Portable Natural Language Interface, Proceedings of A VIGNON: Ninth International Workshop on Expert System and Their Applications*, Avignon, France, May 29-June 2, 1989.

Syu, I. et al, (1990). *A Knowledge-Based Writer's Aid for Simplified English, Proceedings of the Third Florida Artificial Intelligence Research Symposium*, pp. 194-196, April 3-6, 1990.

Inien Syu is a Ph.D. student in computer science at the University of Central Florida. She is currently working on research in the field of knowledge-based information retrieval for her dissertation.

Sheau-Dong Lang is an Associate Professor of Computer Science at the University of Central Florida. His current interests include Database Systems, Real-time simulation and Training, and Knowledge-Based Information Retrieval.