N89-15555

KAM : A TOOL TO SIMPLIFY THE KNOWLEDGE ACQUISITION PROCESS

Gary A. Gettig

Phase Linear Systems, Inc.
9300 Lee Highway
Fairfax, Virginia 22031

ABSTRACT

Analysts, knowledge engineers and information specialists are faced with increasing volumes of time-sensitive data in text form, either as free text or highly structured text records. Rapid access to the relevant data in these sources is essential. However, due to the volume and organization of the contents, and limitations of human memory and association, frequently a) important information is not located in time; b) reams of irrelevant data are searched; and c) interesting or critical associations are missed due to physical or temporal gaps involved in working with large files.

The Knowledge Acquisition Module (KAM) is a microcomputer-based expert system designed to assist knowledge engineers, analysts, and other specialists in extracting useful knowledge from large volumes of digitized text and text-based files. KAM formulates non-explicit, ambiguous, or vague relations, rules, and facts into a manageable and consistent formal code. A library of system rules or heuristics is maintained to control the extraction of rules, relations, assertions, and other patterns from the text. These heuristics can be added, deleted or customized by the user. The user can further control the extraction process with optional topic specifications. This allows the user to cluster extracts based on specified topics.

Because KAM formalizes diverse knowledge, it can be used by a variety of expert systems and automated reasoning applications. KAM can also perform important roles in computer-assisted training and skill development.

Current research efforts include the applicability of neural networks to aid in the extraction process and the conversion of these extracts into standard formats.

INTRODUCTION

As technology advances, reams of information are being put onto electronic media daily. The Federal government alone maintains thousands of databases and there are several agencies such as DTIC and NTIC whose primary mission is to provide information to user communities. With this wealth of information at hand, a tool to access and retrieve relevant information in a comprehensive and timely fashion would greatly enhance productivity. This is especially true when developing an expert system.

Knowledge acquisition is one of the biggest obstacles facing anyone who is attempting to build an expert system. More time is spent on it than any other phase of the expert system life cycle. The first step in knowledge acquisition is domain orientation. This initial gathering of information about a domain is where the Knowledge Acquisition Module (KAM) is the most powerful.

Knowledge engineers generally know little of the subject they are attempting to model. It is often necessary to spend weeks doing background research. This involves reading through many documents and databases, most of which are irrelevant.

During this process, the knowledge engineer may miss valuable information due to time constraints or lack of a structured methodology to locate it. The knowledge engineer may also miss important relationships between objects or ideas. This is caused by temporal relationships within the documentation such that the knowledge engineer can no longer remember previous data and how it relates to what is being read. In addition, spacial gaps in text of related information make it difficult to obtain a comprehensive understanding of a topic or follow lines of reasoning to a conclusion. The above problems are compounded if more than one person is assigned to do the background reading.

The Knowledge Acquisition Module is a word-based natural language processor designed to extract specific knowledge from large volumes of text based files. These files can be in the form of a wordprocessor document, ascii text file, or a highly structured database. Information can reside on disk, in memory, or be ported in via a communication line.

METHODS OF EXTRACTION

A variety of different methods are used to extract information from text. These methods can work at the word, syntax, semantic, or pragmatic level. KAM expounds on two of these methods. The first, context analysis, works at both the word and syntax levels. Words can have a myriad of different meanings when presented alone. Groups of words, however, can have a very specific meaning. This meaning defines an idea or context. For example, while scanning a paragraph the word "BAT" is found. "BAT" can mean several things when viewed by itself. "BAT" could be a noun or a verb, each having a variety of different meanings. Figure 1 gives a sample of some of the different meanings of the word "BAT".

**Figure 1. Stand alone meanings of various words
Adapted from [Webster2]**

BAT (noun) -
1. a solid stick
2. a sharp blow
3. a wooden instrument used for hitting a ball in various games
4. a paddle used in various games
5. rate of speed
6. a flying nocturnal mammal
7. a hag or a witch

BAT (verb) -
1. to strike or hit
2. to discuss at length
3. to wander aimlessly
4. to wink

BASE (noun) -
1. the bottom or lower part of something
2. a main ingredient
3. fundamental part of something
4. the starting place
5. bitter tasting compound
6. the four stations at the corner of a baseball field
7. a center of operations

BASE (verb) -
1. to make, form, or serve as a base for
2. to find a base or basis for (used with on or upon)

BASE (adj) -
1. of little height
2. low in place or position
3. resembling a villain
4. being of low value or inferior properties
5. lacking higher values

DIAMOND (noun) -
1. a native crystalline carbon (gem)
2. a square or rhombus shaped figure
3. something that resembles a diamond
4. a baseball infield

DIAMOND (verb) -
1. to adorn with diamonds

Figure 2. Conceptual Grouping of Battle Adapted from [Webster1]

```
Battle(noun) -
   engagement
   action
   clash -
      brawl, broil, conflict
   assault -
      aggression
         offense -
            aggression, assailment, onslaught
      onfall, onslaught
   attack -
      charge
      drive -
         initiative, push
         raid, blitz, militarization, seizure
      combat -
           action
   contest -
      competition, rivalry, warfare

Battle(verb) -
    war -
       challenge
       struggle -
          endeavor, essay, attempt
       scrimmage -
          affray, skirmish, melee
       assault -
          engage, aggress, beset, storm
       bombard -
          blitz, bomb
          shell -
             bombard, cannonade, rake
          barrage, strike
    fight -
       strive, debate, dispute, resist
          buck -
               dispute, duel, repel, traverse, pulverize,
               unseat, duel
    oppugn -
       contend
```

The next word that might be encountered is "BASE". "BASE" is even more complex than "BAT". As can be seen in Figure 1, "BASE" has different meanings as an adjective, a verb, or a noun. The third word located may be "DIAMOND", which has various meanings as well. But observe what happens when "BAT", "BASE", and "DIAMOND" are put together. The context that the paragraph was alluding to becomes apparent - that of baseball. This was accomplished without baseball being explicitly stated within the document. KAM allows the user to perform this type of analysis through a user defined topical grouping. This grouping can be done at the sentence level, through the use of rules, or at the paragraph level with special topical clusterings. Meanings can also be derived using KAM at the syntactic level by the order and spacial relationships of these words in the topical clusterings.

The second method, conceptual grouping, works in KAM primarily at the word level. It can, however, be modified to include phrases. Conceptual grouping is accomplished by linking with a thesaurus or dictionary of synonyms in order to fully cover the meaning of a specified topic. By specifying one topic and then calling a thesaurus, one will be able to discover links between related ideas that would have otherwise gone undetected by a person searching through large volumes of text. For example, suppose one wanted to discover how a certain document pertains to 'battle'. One could specify 'battle' in the topics list and have KAM help link to related words and concepts. One can even specify the conceptual depth one is required to achieve. Figure 2 shows how easy it is to achieve a high level of conceptual dependency.

KAM uses the thesaurus to break out the user specified topics into parts of speech. It then follows the same part of speech while digging deeper for concepts. For instance, the verb 'battle' chains to the verb 'fight' when looking for related concepts and not to the noun 'fight'. This is to curtail the combinatorial explosion which would result if no constraints were placed on the concept grouping. This can, however, be turned off if necessary. The part of speech of the word can also be specified when giving the topic.

As can be seen from Figure 2, going beyond the third processing level causes two problems to occur. First, the topics become circular and further depth traversal is unnecessary. This can be seen in the case when the noun 'battle' chains to 'assault', which in turn chains to 'offense'. 'Offense' then chains to 'aggression' which was already established by 'assault'. The deeper the level the more circular chaining becomes. The second problem, which is the most difficult to manage, is that a topic tends to veer too heavily from its intended course, resulting in unintended generalizations. The question of when to stop is a matter of how deeply one needs to go to understand or reveal the relationships that are trying to be uncovered. Conceptual grouping is also useful in picking up trends and biases in documents that would have otherwise gone unnoticed.

HOW KAM WORKS

KAM identifies word relationships by their contextual and conceptual dependencies. It will use any text-based or database file to formulate ambiguous or vague relationships, and non-explicit rules or facts into manageable clusters of related information. KAM can also integrate several documents to eliminate spacial gaps between relevant information. The driving force behind KAM is the heuristic file set up by the user. This heuristic file consists of rule forms used in the extraction process. These forms can have associated with them their own set of topics and exceptions. For example, Figure 3 gives a simple heuristic file set up to extract rule forms on the paragraph given below.

> "These instructions pertain to the successful care and operation of the MK-50 automatic weapons under humid, tropical climate conditions. Special precautions may be followed for properly maintaining the lubrication of the cartridge ejection mechanism. If moisture is allowed to build up inside the weapon, jamming may occur during automatic firing. The ammo clip may be removed if jamming does occur. Always be careful to set the safety before attempting to dislodge jammed shells from the barrel. Note that the MK-50 is quite unlike its Uzi counterpart in the fabrication of the bolt and firing pin mechanisms. Therefore, it is important to take precautions in the removal of the firing pin, lest the spring action will come loose."

The extracts produced from this example are presented in Figure 4. Several interesting points can be illustrated from this example. The first is the use of priorities to resolve conflicts that arise in the extraction process. The fact "important" was not extracted because of the higher priority of the rule "comma-lest". Second, the overall generality of the rule forms allows them to be applied generically to many situations. While this example is simple, large heuristic files can be developed and maintained for various applications. In addition, a general heuristic file can be used to extract information from a text file which can, in turn, be used to customize another heuristic file. This is useful when little is known about the particular domain. Thirdly, KAM can use these extractions to aid in constructing rules and facts for direct use by various expert system shells. Below is an example of how KAM incorporates two of the extracts into Goldworks frames.

```
(G-1 (PRINT-NAME ) (DOC-STRING KAM default frame) (IS
KAMFRAME) (SYNTAX RULE) (IF ( moisture is allowed to
build up  inside the weapon)) (THEN ( jamming may occur
during automatic firing)))

(G-2 (PRINT-NAME ) (DOC-STRING KAM default frame) (IS
KAMFRAME) (SYNTAX RULE) (IF ( jamming does occur))
(THEN ( The ammo  clip may be removed)))
```

This example does not imply that all extracts in the raw form are suited for direct placement into an expert system but a good percentage can be incorporated with careful design of the heuristic file.

Another feature of KAM is that it allows the user to perform context analysis and conceptual groupings on paragraphs so that related paragraphs are clustered together. This will eliminate spacial gaps in related information and allow special heuristic files to be designed specifically for these related paragraphs.

Figure 3. Sample Heuristic File

```
RULE if-comma
     IF [1+] , [1+] .
  FORM : IF 1 THEN 2
  PRIORITY : 1
  END

  RULE lone-if
     [1+] IF [1+] .
  FORM : IF 1 THEN 2.
  PRIORITY : 1
  END

  FACT important
     [1+] IMPORTANT [1+]
  FORM : NOTE 1 IMPORTANT 2
  PRIORITY : 3
  END
```

```
RULE comma-lest
    [1+] , [0-0] LEST [1+].
 FORM : IF NOT 1 THEN 3
 PRIORITY : 1
 END

 FACT always
      [1+] ALWAYS [1+] .
 FORM : ALWAYS 2
 PRIORITY : 3
 END
```

Figure 4. Rule Forms : Generated by KAM

1. IF moisture is allowed to build up inside the weapon
   THEN jamming may occur during automatic firing

2. IF jamming does occur
   THEN the ammo clip may be removed

3. IF NOT Therefore, it is important to take precautions
   in the removal of the firing pin
   THEN the spring action will come loose

4. FACT : Always be careful to set the safety before
   attempting to dislodge jammed shells from the
   barrel

RESEARCH TOPICS

There are several areas of research that are currently being addressed. The major thrust is in the area of resolving ambiguous pronouns within the rule forms. This is one of the more difficult problems to solve in natural language processing. The method currently being used by KAM is to allow a toggle between the source text and the extract generated so that the user can resolve any ambiguities. However, if communication lines are used, source reference is impossible. One way to address this problem is to have KAM aid the user in identifying pronoun references so that user would not have to go back to the original source. This aid would come in the form of KAM giving a selection of possible pronoun sources along with a certainty ranking. Current work is being done in this area with neural networks using back-propagation [Allen]. The idea is to have a neural network learn how to identify pronoun references by having it learn from previous examples and generalizing about new instances.

Other areas of interest include how to better control the level of conceptual grouping and how to more efficiently convert the extracts to standard formats such as schemas, frames, and user defined structures. The user defined structures hold the most promise since they allow the user to better control the extraction process. This would allow KAM to perform functions throughout the expert system life cycle.

CONCLUSIONS

KAM provides a simple, robust, and easy-to-use tool for knowledge acquisition. Through the use of context analysis and conceptual grouping, one can cluster paragraphs on selected topics. Specially designed heuristic files can then be used to extract rule forms from the clustered paragraphs. This is usually the most efficient way of approaching the problem of extracting particular information from huge quantities of text. This process can, however, be carried out in reverse.

KAM allows complete control over the extraction and clustering process. In addition to global topics and exceptions, each rule form can have its own special topics and exceptions. This allows the user to better refine the extraction process.

The user can specify to KAM the exact form to use during extraction. The user can also design a template that determines what form the extract will be in. This feature is useful when the extracts are going to be ported into an expert system shell or any other standard format.

The flexibility of KAM allows it to be tailored to just about any application. One of the more interesting ones is training and skill development in a particular domain. Knowledge and skills are obtained much faster when irrelevant information is filtered out.

REFERENCES

Allen, Robert B. "Natural Language and Back-Propagation: Demonstratives, Analogies, Pronoun Reference, and Translation", Proceedings of the First International Conference on Neural Networks, San Diego, June 21-24, IEEE Press, 1987.

[Webster1] "Webster's Collegiate Thesaurus", Merriam-Webster Inc., 1976.

[Webster2] "Webster's Ninth New Collegiate Dictionary", Merriam-Webster Inc., 1988.