The Knowledge Capabilities of the Vocabulary Component of a Medical Expert System

Omar Bouhaddou (1,2), Homer R Warner (1), Hong Yu (1), Michael J Lincoln (1)

(1) University of Utah, Department of Medical Informatics and
(2) Applied Informatics, Inc.
Salt Lake City - Utah

Abstract

In this paper, we describe the ways in which medical knowledge is encoded into the knowledge base dictionary of an expert system (Iliad) designed as a teaching and consulting tool. Starting with a basic hierarchy, attributes have been added to facilitate the entry and display of patient data and inferencing about diagnosis and optimal work-up of the patient. For inferencing, a table of inter-term relationships has been added to the dictionary and techniques for deriving partial information using frequencies of occurrence of parent and child nodes in the hierarchy have been incorporated in the program. The disease independent knowledge included in the dictionary component is necessary to support the expert system's ability to make medically relevant and common sense inferences and generate realistic patient cases.

Introduction

Iliad is a Macintosh-based diagnostic expert system for internal medicine. Iliad provides diagnostic assistance for the major diseases of internal medicine based on a large number of medical relationships included in its knowledge base and its data dictionary. This paper is an attempt to stress the knowledge capabilities of the vocabulary component. These capabilities are crucial in providing the expert system with:

1. the ability to understand the patient findings entered by the user and make its own recommendations and suggestions understood to the user, and
2. the ability to display a measure of common sense and basic medical logic by performing deductions about interdependent observations it knows about.

Relevant information about the disease manifestations is encoded in the dictionary files whenever it is independent of the disease states.

Overview

The Knowledge Base Dictionary (KBD) is an important part of the Iliad expert system knowledge base. The KBD contains textual descriptions and other information about the disease names and findings used to describe patient history, physical examination and laboratory findings. There are currently over 950 diagnostic decision frames supported by 5000 dictionary terms in the Iliad knowledge base. A frame is a computer structure that can be thought of as a form with slots reserved for specific information. A diagnostic frame in Iliad is the collection of disease criteria, the weight of each criteria and the relationships between these criteria necessary to make the diagnosis. Knowledge engineers must have thorough knowledge of the KBD and use that knowledge to assemble disease frames. Iliad users describe patient cases and receive diagnostic consultations based on the interface defined by the KBD. The KBD also contains basic medical knowledge, statistical and deterministic, about interrelationships among medical observations.

Knowledge Base Dictionary Structure

There are six levels of findings in the hierarchical organization of Iliad KBD. The first level is reserved for general categories indicating the source of the information (e.g., medical history, physical examination, radiology). The second level has a different significance depending on the category of information. For instance, the second level in the medical history data class will describe the present history, the previous history, the social history, the family history, and so on. In physical examination data source, the second level defines the organ examined (thyroid exam, eye inspection) or the type of examination performed (e.g., chest auscultation, chest palpation, chest percussion); and in the Lab category, the second level is the name of the test or procedure (e.g., Hematology: CBC, Chemistry: Chem-20, Radiology: CXR). These conventions provide benefits both in terms of easier user interface and inferencing, as will be discussed later in the paper.

Meaningful findings can be formed by beginning at the second level and including zero or more lower level terms. Clinical terms (e.g., present history of cough) start at the third level. Fourth, fifth, and sixth level findings are specific modifications of their "parents." Any lower level findings is expressed in such a way that it can be concatenated with its higher level parents to create a complete, English-like, more specific finding. For example, "cough" is the third level finding, "productive of sputum" is the fourth level term, and "that is muco-purulent " is a fifth level term in this hierarchy. From the user perspective, the complete fifth level finding becomes "history of cough productive of sputum that is muco-purulent". The hierarchical organization of the dictionary into single concept per level (i.e., atomic structure) provides several benefits. First, it enhances patient data entry and review by allowing medically relevant groupings and clear formatting in the listing of the patient record. Second, the organization provides a basis for powerful inferencing among findings in disease frames which use terms at different levels of such a hierarchy. Below are excerpts from Iliad current dictionary that illustrate the hierarchical relationships between terms:

- Medical History
  - Present history:
    - chest pain
      - at rest
      - [with a specific time pattern]
        - that recurs
        - that began recently
        - occurring at the same time each day
        - that came on suddenly
        - increasing in frequency or duration
        - worse at night
        - persistent (not intermittent)
        - lasting === minutes
      - [with a specific quality]
        - that is sharp or stabbing
        - that is burning
        - that is crushing
        - that is aching
        - increasing in severity
        - dull and unilateral
      - [with a specific location]
        - under the sternum
      - radiating
        - to the shoulder, arm or jaw
      - [with specific aggravating factors]
        - breathing deeply
        - coughing
        - lying flat
        - eating
        - bending or stooping
        - stress
        - moving about or changing position
        - [with exercise]
          - with less exercise than in the recent past
      - [with specific relieving factors]
        - rest
        - nitroglycerin
          - nitroglycerine completely
        - belching
        - milk or antacids
          - in a few minutes
  - Past medical history
  - Family medical history
  - Social history
  - Medication history

- Physical Examination
  - Otoscopy:
    - abnormal auditory canal
      - that is red or swollen
      - purulent drainage
    - abnormal tympanic membrane
      - that is red or dull
      - that appears inflamed
      - that is retracted or bulging
      - with loss of the light reflex
      - with perforation
      - bullous myringitis
      - middle ear fluid
    - abnormal pinnae

- Radiology
  - Chest x-ray
    - [shows bony abnormalities]
      - shows abnormal ribs
        - due to notching
        - with destruction/erosion
        - due to tumor
        - with destruction in the first three ribs
      - shows abnormal sternum
        - due to tumor
      - shows abnormal clavicle
        - due to tumor
        - with destruction of the lower cervical
          - or upper thoracic vertebra
      - shows abnormal scapula
        - due to tumor
    - [shows parenchymal abnormalities]

Figure 1: In the dictionary, the terms which define patient symptoms and diseases are organized in parent-child relationships with more general information branching to more specific information at the next level of the tree. Above are samples of the ILIAD dictionary from the Medical History, Physical Examination and X-ray findings sections of the dictionary. Phrases in brackets [ ] will not be indicated or printed by ILIAD if they have modifiers at a lower level. If the phrase in brackets has no additional modifiers, the phrase in brackets will be displayed or printed, but without the [ ]. The === sign, when used in the dictionary, are simply place holders for numerical values and are replaced by actual patient information.

The vocabulary component of the Iliad expert system is made of several files. The textual descriptions of the dictionary terms are stored in the text file and each finding has a corresponding hierarchical code which is stored in the code file. Another file, the keyword file, contains keywords extracted from the textual description of the findings and additional keywords added manually (i.e., synonyms, abbreviations). A KBD editor program allows a user to search, add or change any attribute of a finding. Usually, a single keyword will retrieve a wide range of findings (e.g., keyword "pain" will bring back chest pain, abdominal pain, joint pain, tenderness to palpation, ...). Adding qualifier keywords provides the user with more specific results. Figure 1 from the KBD editor shows the entry form for a dictionary item. Also stored with each dictionary entry is the type of value or status allowable for the entry.

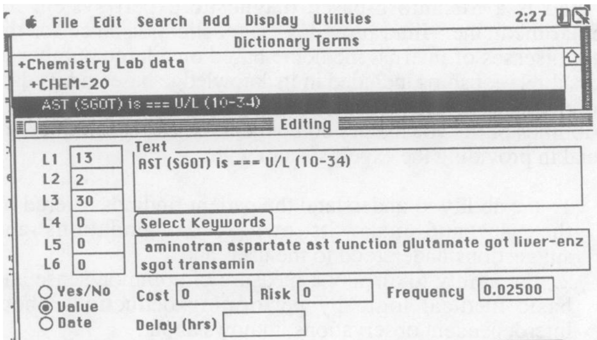

Figure 2: The entry form for dictionary items in the Iliad system. L1 to L6 indicate the 6 level hierarchical address of the node. Note that the cost of SGOT is set to 0 because only the parent term, Chem-20, is charged. The slot delays (hrs) is not used at the present time, but will indicate the delay between order and results.

The status of a dictionary item can be a numerical value (e.g., urine volume measurement or WBC count) or a YES / NO / UNKNOWN status (e.g., present history of chest pain or sputum culture positive for pneumococci).

Additional attributes which are stored in the KBD help the program emulate more closely the natural clinical environment. One such attributes is the cost of acquiring each finding. An arbitrary cost of $1 for each historical finding, $2 for each physical exam finding and the actual cost (charge) for each lab test and diagnostic procedure. The program uses the cost information to suggest the most cost-effective workup strategy. This allows Iliad to typically pursue history questions first, then physical examinations and less expensive lab tests before choosing expensive procedures.

The frequency with which each finding occurs in a hospital population is stored in the KBD. The acquisition procedures for these statistics have been presented in 1989, at the last SCAMC meeting [3].

The KBD also flags whether a historical finding can be a patient's chief complaint and whether a decision frame is a final diagnosis. Patients always come to hospital with a chief complaint, and Iliad patient case simulator tries to generate "artificial" patient cases resembling as closely as possible real cases. In a simulation session, Iliad begins by presenting the user with the patient's chief complaint and then lets the user ask for further questions to workup the case. To meet this need, appropriate historical findings are assigned the chief complaint flag (e.g., cough, chest pain, fever) while other historical items are not (e.g., exposure to uranium dust, fingers demonstrate hyperemia during recovery). There are two kinds of decision frames in Iliad knowledge base: final diagnoses (e.g., acute myocardial infarction, pulmonary embolus) and intermediate decisions or "clusters" (e.g., lung consolidation, risk factors for AIDS). Most of the final diagnosis frames are represented with a Bayesian model. A Bayesian model refers to a disease description that includes the disease prevalence, a list of disease manifestations weighted by their sensitivity and specificity and the use of Bayes conditional probability theorem to update the likelihood of the disease given the information at hand. We assign a special flag to cluster frames in order to prevent from including them in Iliad differential diagnostic list.

<u>The Word-relation file</u>

User may enter different words to search for the same term because there are many synonyms, abbreviations and other relations in medical vocabulary. For example, white blood cells can be searched by its abbreviation "WBCs" or by the synonym "leukocytes." Some users may type in "leucocytes" instead of "leukocytes," and Iliad recognizes these variations. These commonly accepted variations are stored in a separate word-relations file. The file also allows a search for related concepts, which are not necessarily synonyms or abbreviations. For instance, using the keyword "time" in association with a symptom name will match terms indexed under "morning", "afternoon", "prodrome", or "recent." The complete relation table for the concept time in Iliad is shown in figure 3.

time duration length when onset begin began stop ceased
- sudden abrupt acute rapid
- slow insidious chronic gradual step-wise stepwise step-wize
- morning
- afternoon
- evening
- after post subsequent
- before pre previous prior
- during
- last
- post-prandial postprandial
- preprandial pre_prandial
- intermittently periodically
- night nocturnal
- day
- prodrome
- sustain prolong persist persistent long
- short
- past
- current
- recent

Figure 3: A sample relation for the concept "time" from the Iliad word-relation file. Synonyms are listed at the same level while specific instances of the main concept are listed in indentation.

Relations have the risk to decrease the precision of keyword queries because of unanticipated interactions with the phraseology of the KBD terms. But, we have found that carefully constructed relations increase the recall of keyword searches with little loss in precision, especially if multiple keywords are used [4].

<u>The data-relation file</u>

The data-relation file embodies interrelationships among the medical concepts contained in the knowledge base. It provides the program with a measure of common sense and logic. For example, if a patient blood type is O then it is not A nor B nor AB; or if the patient is female then the hypothesis of 'Testicular Failure' is not applicable.

A simple hierarchical organization of diagnostic concepts is able to encode a parent-child relationship but fails to detail interdependence between siblings or distant nodes, and does not allow multiple contexts for the same term. Some of these relationships can be captured in the knowledge base when they lead to higher interpretations [5]. For instance, chills, fever and sweat can all be the expression of a common cause and be interpreted as "systemic signs of infection." But, the aim of the data-relation file is to infer the status of an observation based on logical deductions from the presence or absence of other observations. It encourages the system to display a common sense relieving the user from entering information which can be inferred from information already known to Iliad.

<u>The user interface capabilities</u>

This user interface refers to the ability of the user to tell things to the system as well as understand what the system tells him or her. The user interface is a fundamental factor in determining the success of a medical expert system.

<u>Understanding prompts from the system</u>

The hierarchical organization of the dictionary terms is reflected in the way findings are presented to the user for an answer. For example, after answering "yes" to present history of chest pain, Iliad will ask about specific chest pain qualifiers

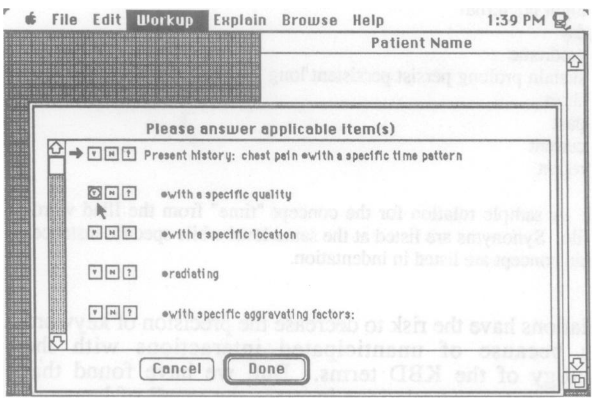

Figure 4: After answering yes to chest pain the user is prompted, in an indented display, to enter specific chest pain qualifiers present in his/her patient.

Display of the patient record

Several key features of the patient record window (see figure 5) that help a user easily and quickly "process" its contents depend on knowledge stored in the KBD [6]:

1) The organization of the findings retains the dictionary hierarchical structure of findings, but is not restricted by the classification of the data dictionary. It is also independent of the order the findings are entered.
2) The display allows the hierarchical relationships among different findings to be displayed to the user as an integral part of the patient record. The hierarchical structure of the Iliad dictionary is explicitly represented in this patient record presentation by indentation.
3) Positive findings are emphasized in bold
4) Negative findings are re-stated in negative terminology
5) Abnormal test values are identified and displayed in bold.

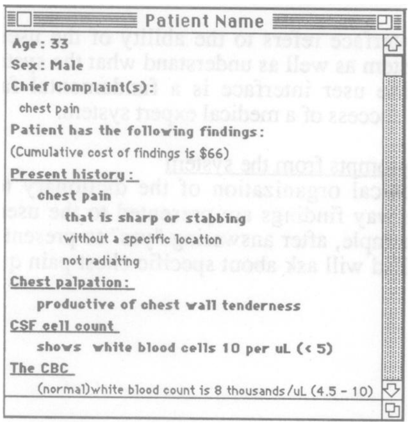

Figure 5: An example of a patient data window display in Iliad. Categories are separated and within each category related findings are showed in indentation.

Medical inferencing capabilities of the dictionary

The information stored in the dictionary files allows for several types of inferencing. Some are derived from common sense knowledge while others are more quantitative deductions. We believe, that a good fraction of knowledge and "smartness" displayed by an expert system can be attributed to the structure of the vocabulary component.

In this section, we will review the inference capabilities of these structures.

Inferencing capabilities of a hierarchical dictionary

A first set of inferences are derived from the parent-child relationships inherent to the hierarchical organization.

- IF Parent is absent THEN the children are absent

This relation is exemplified by the deduction from NO cough to NO cough productive of sputum to NO hemoptysis. If the answer is YES to the parent level "present history of cough" then Iliad may ask further questions about the children levels (e.g., present history of cough productive of sputum), otherwise no further question is asked and the status of all children levels is set to NO.

- IF child is present THEN parent is present

For example, if we know there is a RUQ abdominal pain then we can infer abdominal pain is present.

- IF parent is present THEN the probability that the probability of any one child might being present is increased

Indeed, based on the conditional probability calculation, the probability of presence of a child is higher if we know the parent is present than if we know that the status of the parent is absent or unknown. Let us illustrates the computations associated with this type of quantitative inferencing. As described above, each dictionary term has a frequency of occurrence in the general population (i.e., hospital population). For instance, pneumonia has a prevalence of 0.025 and the symptoms cough and productive cough are estimated to occur in 10% and 5% of the hospital patients respectively. In patients with pneumonia, productive cough occurs in 80% of the cases. In Iliad's Bayesian model, the posterior probability of pneumonia can be obtained using the general Bayes formula:

$$Pd/s = (Pd * Ps/d) \div Ps = 0.025 * 0.80 \div 0.05 = 0.40$$

where:
Pd is the a prior probability of pneumonia (= 0.025),
Pd/s is the posterior probability of pneumonia given productive cough is present
Ps/d is the probability of a productive cough in pneumonia patients (=0.80), and
Ps represents the frequency of productive cough across all diseases (=.05)

Now, if all we know is that the patient coughs, this should update by some fraction of belief our consideration of pneumonia. This is done by substituting the frequency of cough for the frequency of productive cough:

$$Pd/cough = 0.025 * 0.80 / 0.10 = 0.20$$

This amounts to a smaller probability increase for the diagnosis of pneumonia than if productive cough was actually observed, but a higher probability than if cough was unknown (i.e., Pd/s = Pd = 0.025) or if cough was absent (Pd/no cough productive = 0.025 * 0.20 / 0.90 = .0056).

- Frequencies provide a prevalence for deterministic frames

A Boolean frame is a grouping of disease manifestations related with a Boolean logic statement (i.e., IF-THEN with ANDs and ORs). An example of a Boolean frame is: if 'PE shows cyanosis OR pO2 < 60 OR O2 saturation < 90' then

'Hypoxemia'. The frequency of each finding in a Boolean frame is used effectively to provide a measure of how close the decision is to being true or to being false. This is done by dividing each of the weights derived from the Boolean statement by its frequency and then normalizing [7].

Inferencing from knowledge in the data-relation file

The data-relation file allows inferencing among dictionary terms. Several types of data relations have been considered so far. Figure 6 shows examples of data-relations.

Not Applicable	
1.1.8.1.0.0	male
1.10.16.6.8.0	exacerbated by menstrual period
1.10.156.24.0	GYN bleeding
1.10.200.0.0.0	abnormal menses (if female)
1.10.202.0.0.0	abnormal vaginal discharge
1.10.204.0.0.0	perineal pruritis
1.10.224.0.0.0	pregnancy (if female)
1.11.88.0.0.0	galactorrhea
Mutually exclusive	
24.63.2.0.0.0	thyroid biopsy shows follicular carcinoma
24.63.4.0.0.0	thyroid biopsy shows papillary carcinoma
24.63.6.0.0.0	thyroid biopsy shows medullary carcinoma
24.63.8.0.0.0	thyroid biopsy shows anaplastic carcinoma
Same Value As	
13.1.12.0.0.0	chem-7, creatinine is === mg/dL (0.8-1.4)
13.2.18.0.0.0	chem-20, creatinine is === mg/dL (0.8-1.4)
Equivalent	
1.10.25.2.0.0	hx of vomiting: hematemesis
1.10.77.10.0.0	hx of blood loss: hematemesis
In sequence	
15.40.0.0.0.0	Histology of bone marrow aspiration
18.2.2.0.0.0	Philadelphia chromosome present

Figure 6: Examples of a data relations from Iliad knowledge base dictionary.

The data-relation file also serves as a master item table that supports multiple occurrences of an item in the dictionary hierarchy (i.e., tangled hierarchy). For instance, the finding hematemesis exists as modifier of history of vomiting and a modifier of blood loss history, and whenever one occurrence is set to true or false, through the data-relation file, the other is set to the same status. However, only the user-entered item is stored in the patient chart (say vomiting blood), the related items are only showed when necessary (e.g., as an explanation for iron deficiency anemia) therefore permitting to upgrade of the data-relation file without changes to the original patient records.

Conclusion

In summary, we find that more and more knowledge is being encoded in the dictionary component of the Iliad program. This knowledge is disease-independent and represents the relationships among terms that are used to describe one or more diseases. In many cases these relationships are needed for the expert system to display appropriate common sense medical behavior. Starting with an unambiguous hierarchy (i.e., each term has a unique code to represent its hierarchical position), new attributes and inter- relationships between findings have been added to enhance the knowledge capabilities of the vocabulary component. The new attributes include:

- additional synonyms and abbreviations
- status (numerical value vs present/absent)
- actual cost
- normal values in the case of laboratory results
- specific flags (e.g., chief complaint, final diagnosis)
- frequency of occurrence of the finding in the general hospital population

Validation of knowledge-based simulated cases has been compared to the "Turing test" for decision modeling in an expert system. We argue that an important step toward that goal is in defining and validating the knowledge capabilities of the vocabulary component. Our growing experience with the simulator mode of Iliad dictated most of our needs for enhanced vocabulary structures. When simulating realistic patient cases, the simulator needs to know the statistical relationships between a disease and the disease manifestations, and the meaningful inferences that should be made on the basis of the simulated findings. Furthermore, the simulator needs to know the allowable descriptors attached to each diagnostic concept to insure more complete history narrations. In the future, we contemplate making simulations even more realistic by adding images and sounds to appropriate dictionary terms. We have already experimented with adding pictorial representations to diagnostic concepts in one domain of Iliad knowledge base using a generic videodisc [8].

This work was supported in part by grant number 5 RO1LM04604-03 from the National Library of Medicine.

References

[1] Warner HR, Haug PJ, Bouhaddou O, Lincoln MJ, Warner HR, Jr., Sorenson D, Williamson JW, Fan C. ILIAD: an Expert Consultant to Teach Differential Diagnosis. Proceedings of the 12th Symposium on Computer Applications in Medical Care (SCAMC), Robert Greenes ed., IEEE Computer Society Press 1988:371-374.

[2] Cundick R, Turner CW, Lincoln MJ, Buchanan JP, Anderson C, Warner HR, Bouhaddou O. ILIAD as a Patient Case Simulator to Teach Medical Problem Solving. Proceedings of the 13th Symposium on Computer Applications in Medical Care (SCAMC), Lawrence C. Kingsland III ed., IEEE Computer Society Press 1989:902-906.

[3] Yu H, Haug PJ, Turner CW. Embedding Clinical Statistics in a Data Dictionary for Use in an Expert System. Proceedings of the 13th Symposium on Computer Applications in Medical Care (SCAMC), Lawrence C. Kingsland III ed., IEEE Computer Society Press 1909:235-241

[4] Fan C. Odyssey: A Computer Program to Access Medical Knowledge. A thesis for the Master's degree in Medical Informatics at the University of Utah, August, 1987.

[5] Lincoln MJ, Turner C, Hesse B, Miller R. A Comparison of Clustered Knowledge Structures in Iliad and in Quick Medical Reference. Proceedings of the 12th annual Symposium on Computer Applications in Medical Care (SCAMC), Robert Greenes ed., IEEE Computer Society Press 1988:131-135

[6] Sun Q. Patient Record Presentation in a Hierarchically Based Medical Expert System. A thesis for the Master's degree in Medical Informatics at the University of Utah, August, 1989.

[7]. Sorenson DK, Cundick RM, Fan C, Warner HR. Passing Partial Information among Bayesian and Boolean Frames. Proceedings of the 13th Symposium on Computer Applications in Medical Care (SCAMC), Lawernce C. Kingsland III ed., IEEE Computer Society Press 1989: 50.

[8]. Stensass SS, Bouhaddou O, Hardy LH, Sorenson DK, Dougherty NE, Altman SJ. The educational Potential of Linking a Videodisk and an Expert System. Proceedings of the 13th Symposium on Computer Applications in Medical Care (SCAMC), Lawernce C. Kingsland III ed., IEEE Computer Society Press 1989: 864-868.