International Journal of Artificial Intelligence in Education 19(2009) 309-334
IOS Press

Expanding the Space of Plausible Solutions in a Medical Tutoring System for Problem-Based Learning

Hameedullah Kazi, *Computer Science & Information Management Program, Asian Institute of Technology, Pathumthani 12120, Thailand*
hameedullah.kazi@ait.ac.th
Department of Computer Science, Isra University, Hala Road, Hyderabad, Pakistan
hkazi@isra.edu.pk

Peter Haddawy, *Computer Science & Information Management Program, Asian Institute of Technology, Pathumthani 12120, Thailand*
haddawy@ait.ac.th

Siriwan Suebnukarn, *School of Dentistry, Thammasat University, Pathumthani 12121, Thailand*
ssiriwan@tu.ac.th

Abstract. In well-defined domains such as Physics, Mathematics, and Chemistry, solutions to a posed problem can objectively be classified as correct or incorrect. In ill-defined domains such as medicine, the classification of solutions to a patient problem as correct or incorrect is much more complex. Typical tutoring systems accept only a small set of approved solutions for each problem scenario fed to the system. Plausible student solutions that fall outside the scope of this small set of approved solutions are rejected as being incorrect, even though these solutions may be acceptable or close to acceptable. This leads to brittleness in the evaluation of student solutions. This paper describes a tutoring system for medical problem-based learning (PBL), which can accept a wide variety of plausible solutions without placing an extensive burden on knowledge acquisition. A widely available medical knowledge source is deployed as a domain ontology, and concept relationships in the ontology are used to make inferences and expand the space of plausible solutions beyond the scope of solutions explicitly provided to the system. Parent-child relationships are used to infer generalized solutions, whereas relationships of synonymy are used to infer alternate solutions. Evaluations of the system after expanding the solution space indicate accuracy close to that of human experts, who agreed among themselves with Pearson Correlation Coefficient of 0.48 and $p < 0.05$. The system precision dropped by 32%, while the recall increased by five times. The geometric mean of sensitivity and specificity was increased by 0.33.

Keywords. Robustness, ill-defined domains, medical PBL, UMLS, knowledge acquisition, ITS.

INTRODUCTION

In well-defined domains such as Physics, Mathematics, and Chemistry, correct solutions to a given problem fall within a relatively small range of possible solutions. In such domains, solutions to a posed problem can easily and objectively be classified as correct or incorrect. In ill-defined domains

such as medicine, the classification of solutions to a patient problem as correct or incorrect is much more complex. Solutions can vary in quality along a number of dimensions such as comprehensiveness and level of detail. In addition, a large number of solutions are possible depending on the perspective from which one chooses to analyze and solve the posed problem, particularly in ill-defined domains such as medical reasoning (Pople, 1982).

Tutoring systems typically contain a set of problem scenarios and a set of solutions approved by domain experts. These approved solutions are either explicitly stored in the system or are generated by applying rules to a knowledge base. Student solutions that match any of the approved solutions for that particular problem scenario are considered acceptable, whereas plausible student solutions that fall outside the scope of this set of approved solutions are rejected as being incorrect, even though these solutions may be acceptable or close to acceptable. This forces students to memorize expert solutions and can limit student creativity. Being able to accept a broader set of solutions provides an approach to learning that promotes free thinking and novel solutions. Students would be able to use their understanding of concepts and concept relationships and learn how to apply their knowledge to given problems. This is especially relevant for the domain of medical problem-based learning (PBL), where a diverse set of solutions may be acceptable. The tutoring systems' rejection of plausible student solutions is quite contradictory to how a human tutor is expected to behave and leads to system brittleness in the evaluation of student solutions. In order for a tutoring system to exhibit robust human-level tutoring, it needs broad knowledge to allow students to explore a large space of solutions and work creatively. A tutoring system should also provide the student with a broad scope of solution representation where the student is able to form or select solution elements from a large repository of domain concepts. Explicitly encoding the complete range of possible solutions into a tutoring system is a daunting task and demands extensive knowledge acquisition from domain experts, which may not be feasible.

Authoring tutoring systems typically requires knowledge acquisition in the three areas of domain expert knowledge, student model and pedagogical model (Murray, 1999). Domain expert knowledge is acquired to equip the system with the curriculum knowledge that is to be taught to the students; pedagogical knowledge is acquired to equip the system with teaching techniques and strategies; and knowledge of the student model is acquired to help the system assess the knowledge level of the student so the system can provide hints based on the student's knowledge level. Acquiring and encoding the relevant knowledge can lead to a large overhead in the development time of a tutoring system (Anderson, Corbett, Koedinger & Pelletier, 1996; Mitrovic, 1998). It makes knowledge acquisition all the more challenging, if the tutoring system is expected to be robust in its evaluation of student solutions as discussed earlier. Manually encoding all knowledge into the system so that it can accept the full range of plausible solutions is not feasible and places great burden on human domain experts whose time is very costly. A natural choice to overcome this knowledge acquisition bottleneck is to make use of existing knowledge available for reuse and sharing. The use of ontologies is a viable alternative in reducing the burden of knowledge acquisition for knowledge based systems.

Specialized ontologies have been employed in the design of various tutoring systems (Crowley & Medvedeva, 2006; Day et al., 2005; Lee, Sen & Evans, 2002; Oguejiofor, Kicinger, Popovici, Archiszewski & Jong, 2004), which often required cumbersome encoding of the ontology through the help of domain experts. Oguejiofor et al. (2004) employed a customized ontology in the design of a system for teaching the domain of personal air vehicles. Their ontology contained names of parts and systems related to the operation of an air vehicle and the relationships between those parts and systems. Crowley and Medvedeva (2006) designed an ontology for a system that teaches

dermatopathology. Their ontology is comprised of concepts that describe diseases and symptoms and the relationships between respective diseases and symptoms. Typical encoding of the ontology in these systems comprises only the concepts relevant to the particular problem scenarios that are to be fed to the tutoring system. Thus, the addition of new problem scenarios to the tutoring system requires additional encoding of domain concepts and their inter-relationships. The Constraint Acquisition System (Suraweera, Mitrovic & Martin, 2005) used a more automated approach for encoding the ontology constraints by learning from examples using constraint-based modeling. However, it still required the initial design of the ontology to be defined manually.

In the next few sections, we describe how the task domain of medical PBL relates to some of the characteristics pertaining to ill-defined domains and how our system design addresses the issues related to those characteristics. We describe how the burden of knowledge acquisition can be reduced for a medical tutoring system through the use of the broad and freely available medical knowledge source the Unified Medical Language System (UMLS), distributed by the U.S. National Library of Medicine (NLM) (U.S. National Library of Medicine, 2008). We also describe a mechanism for exploiting the information structure in UMLS, through which the tutoring system can accept a range of plausible solutions larger than the explicitly encoded scope of solutions.

RELATED WORK

This paper discusses work that is relevant to a number of issues. Expanding the space of plausible solutions relates it to the issue of robustness in intelligent tutoring systems and tackling the burden of knowledge acquisition. Furthermore, the task of medical PBL is in many ways similar to other ill-defined domains as discussed later. At the same time, this work is also strongly tied to the design of medical tutoring systems and the use of UMLS in intelligent systems.

Robustness and Knowledge Acquisition

The issue of brittleness and the burden of knowledge acquisition has been addressed in the design of various intelligent tutoring systems (Kumar, 2002; Remolina, Ramachandran, Ru, Stottler & Howse, 2004; Rubin, Rush, Smith, Murthy & Trajkovic, 2002). Kumar (2002) discusses the use of model-based reasoning for domain modeling in the context of a web-based intelligent tutoring system for helping students to learn to debug C++ programs. Arguing that rule-based systems are not flexible and are not adaptable to varying system behavioral discrepancies, he proposes model-based reasoning as an alternative.

The KASER (Rubin et al., 2002) design is implemented in a tutoring system that teaches the science of crystal-laser design. They argue that production rules, when found to be in error, are corrected through explicitly encoding the revised rule back into the system, leading to burdensome knowledge acquisition. They propose a production system, which initially requires explicit encoding of rules and can later self generate other rules as extensions to the ones created earlier, easing the burden of knowledge acquisition. The brittleness they address is related to the limitation of sufficient encoding of rules that form the system's domain knowledge.

In the design of an intelligent, simulation-based tutor for flight training, Remolina et al. (2004) address brittleness that relates to the issue of student modelling. They discuss their earlier work on a rule-based approach found to be brittle in the instructional approach that was encoded deep into the

logic of the system. Their improved design led to the Adaptive Instructional System, which provided student modelling of personality traits so the tutor could adapt to different student needs. The type of brittleness we address in this paper relates to the limitation of system domain knowledge in evaluating student solutions.

Intelligent Tutoring Systems in Ill-Defined Domains

Work on Intelligent Tutoring Systems in ill-defined domains has covered various domains such as law (Aleven, Pinkwart, Lynch & Ashley, 2006) and medicine (Gauthier, Lajoie & Richard, 2007). Easterday, Aleven and Scheines (2007) studied how causal graph diagrams could be used in an ill-defined domain such as analyzing policy arguments. They conducted experiments with participants who were required to answer policy questions through text or using a combination of both text and diagrams.

Le and Menzel (2007) presented a constraint-based modelling approach to describing the solution space for a Prolog logic programming task. They claimed that despite the fact that the task of solving a Prolog programming problem had well-defined start and goal states, and that the solutions were also verifiable, the task of logic programming is still ill-defined. They argue that the task is essentially a design problem where the code can be made clearer, more maintainable and reusable, all of which are subjective attributes, hence categorizing the task as ill-defined. They employed constraints to specify the requirements of the task at hand and evaluate solutions based on the satisfaction of the constraints.

The Rashi system (Dragon & Woolf, 2006) adopts an inquiry-based approach to learning in the four ill-defined domains of geology, biology, art and history. This work is based on providing feedback that stimulates critical thinking among students and requires them to present appropriate evidence for their hypotheses. Student solutions are assessed and feedback is provided using a set of expert rules and an expert knowledge base.

Medical Tutoring Systems

The designs of medical tutoring systems built to date have typically been based on customized knowledge bases that offer students a limited set of medical terms and concepts with which to form their solution. The CIRCSIM-Tutor (Mills, Evens & Freedman, 2004) teaches cardiovascular physiology by describing a perturbation of a cardiovascular condition. The system initiates a Socratic-style question answer dialog with the student to help the student in reasoning towards the correct solution. The system design places emphasis on qualitative reasoning, but the scope of hypothesis (solution) representation is narrow, as students are confined to assigning values to a small set of variables for forming their hypothesis.

The SlideTutor (Crowley & Medvedeva, 2006) teaches students dermatopathology by presenting a visual slide as a problem scenario and asking students to classify the diseases. After observing the visual evidence presented in the slide, students present their hypothesis through a mouse driven menu selection, identifying features and their attributes from an ontology that has been manually encoded for the problem scenarios fed to the tutoring system. Solutions accepted by the tutoring system are also based on the ontology customized for the tutoring system. The system cannot accomodate alternative plausible hypotheses that may lie beyond the scope of this customized ontology.

UMLS in Intelligent Systems

The UMLS has been studied for use in many intelligent system applications. Achour, Dojat, Rieux, Bierling and Lepage (2001) describe a knowledge acquisition tool and how it could be used to use and share knowledge from UMLS. Their work is primarily based on providing knowledge bases for clinical decision support systems. Their focus is not on using the semantic types and concept relationships in UMLS for reasoning purposes, but on using UMLS knowledge sources as a repository of terms from which a domain ontology could be easily constructed. Crowley, Tseytlin and Jukic (2005) describe ReportTutor, a tutoring system that presents students with a visual slide for inspection and a natural language interface for typing their diagnostic report. They employ the UMLS MMTx to match concepts in the report to concepts in the NCI Metathesaurus (National Cancer Institute, 2009) and validate the report findings against a domain ontology.

UMLS has also been studied for purposes of knowledge extraction by examining the relationships among medical concepts found in medical documents. Burgun and Bodenreider (2001) studied the relationships among Metathesaurus concepts co-occurring in Medline citations. Mendonca & Cimino (2000) described their work on extracting knowledge from MEDLINE citations for purposes of building a knowledge base. They analyzed the search results to determine which semantic types were relevant to what kinds of questions in Evidence Based Medicine, such as diagnosis, etiology, therapy or prognosis. Srinivasan, Rindflesch, Hole, Aronson and Mork (2002) evaluated the extent to which Metathesaurus concepts were found in the full collection of Medline citations. Their results revealed that 30% of concepts were found in the titles and abstracts of articles in the literature.

UMLS has also been employed as a thesaurus to support query expansion, a process of enhancing information retrieval by reformulating the original query. Hersh, Price and Donohoe (2000) assessed query expansion based on a MedLine test collection (OHSUMED) of queries. They expanded the query base by employing synonymous and hierarchical relationships between concepts and related term information and concept definitions in the UMLS Metathesaurus, and reported the precision and recall on the retrieved results. Liu and Chu (2007) proposed a method for expanding queries for purposes of medical text retrieval, based on the query scenario. They employed the UMLS knowledge sources to ascertain the co-occurrence of concepts in medical documents and exploited the UMLS semantic network relationships between the semantic types of co-occurring concepts to further refine the search. In order to expand the query base, they also employed the hierarchical relationships between concepts to add terms to the original query.

To the best of our knowledge, UMLS has not been used as the main knowledge source for inference purposes in an intelligent tutoring system. The Docs ‘n Drugs tutoring system (Martens, Bernauer, Illmann & Seitz, 2001) uses medical terminologies that are a subset of UMLS to allow students to choose concepts from these incrementally expandable terminologies. However, this system does not exploit the knowledge structure within these terminologies.

MEDICAL PBL & ILL-DEFINEDNESS

A PBL session in the medical domain typically comprises a group of 6-8 students who are given about two hours to solve a given problem. The dynamics of a PBL session require close tutor attention to a small group of students and place high demands on costly medical faculty time. Intelligent tutoring

systems may offer a cost effective alternative to help students acquire the required clinical reasoning skills (Suebnukarn & Haddawy, 2007).

Based on the problem scenario presented, students form their hypotheses in the form of causal graphs, where graph nodes represent hypothesis concepts and directed edges (causal links) represent cause-effect relationships between respective concepts. The hypothesis graph is based on the Illness Script, where hypothesis nodes may represent enabling conditions, faults or consequences (Feltovich & Barrows, 1984). Enabling conditions are factors that trigger the onset of a medical condition (e.g., aging, car accidents, or smoking); faults are the bodily malfunctions that result in various signs and symptoms (e.g., myocardial infarction, diabetes); consequences are the signs and symptoms that occur as a result of diseases or disorders (e.g., fatigue, coughing, swelling). A solution is considered complete when the students have elaborated all plausible causal paths leading from enabling condition nodes to relevant fault nodes and from the fault nodes to the relevant consequences nodes.

Lynch, Ashley, Aleven and Pinkwart (2006) identify five key characteristics of ill-defined domains related to *verifiability*, *formal theories*, *task structure*, *open-textured concepts* and *overlapping sub-problems*. The task domain of medical PBL bears some of these characteristics relating to *verifiability*, *design task structure* and *open-textured concepts*. Depending on the perspective from which one chooses to analyze, the classification of a solution as correct or incorrect can be quite complex and difficult to verify. The task structure of forming a causal graph that explains the phenomenon described in the patient problem requires creativity in analysis and identification of the various medical concepts that could potentially lead to the prevailing medical condition. Thus, the task structure of forming a solution for a problem in medical PBL is similar to a design problem. Additionally, since alternate acceptable solutions differ in the level of detail with which they are presented, open-textured and generic concepts can be described instead of specific terms (e.g. the concept *infectious disease of the lung* may replace *pneumonia* in the formation of a medical PBL hypothesis).

In order to delve deeper into the dynamics of the ill-definedness of medical PBL, we conducted exclusive interviews with three medical experts from Isra University. Their responses revealed that because of the ill-defined nature of the medical PBL task there could be multiple plausible solutions to a problem and also a fair bit of disagreement among medical experts regarding their acceptability. The difference of opinion among medical experts or variation in multiple acceptable solutions to a problem could occur due to a variety of reasons discussed below. Some of these factors also coincide with the characteristics identified by Jonassen (1997) and Pople (1982) that make medical diagnostics an ill-structured problem.

Variation in multiple acceptable solutions can arise as a result of differential diagnosis: a process of comparing the possibility of different diseases that lead to similar symptoms. For example, a patient reporting abdominal pain may be considered as possibly having peptic ulcer, gall bladder disorder, or pancreas disorder. Upon a physical examination of the abdomen that reveals tenderness, a physician may eliminate the possibility of peptic ulcer and weigh the chances of the remaining two causes. Thus, prior to the emergence of further evidence, two solutions may differ over the actual cause of the patient symptoms.

Difference of opinion may also occur due to the expertise of the PBL tutor or facilitator. This has also been pointed out by Das, Mpofu, Hasan and Stewart (2002) while reporting an evaluation of PBL tutors by students. A cardiologist would be more inclined to diagnose a patient as having heart disease compared to a physician who specializes in medicine or gastro-enterology. Furthermore, a PBL tutor

having advanced knowledge in physiology would be more accepting of problem solutions that address disorders of the endocrine system compared to a tutor with lesser knowledge in the same area.

Differences may also arise due to the choice of alternate or synonymous terms. For example, one solution might present Coronary Atherosclerosis as a cause of Myocardial Infarction, while another solution may list Coronary Atheroma as a cause . Both are plausible solutions because Coronary Atheroma can serve as an alternate term for Coronary Atherosclerosis.

Solutions can also vary on the basis of terms being more general or specific. For example, one solution might show Glucose Metabolism Disorder as a cause of Hyperglycemia, while another may show Metabolic Diseases as a cause. Here Metabolic Diseases is a general concept and Glucose Metabolism Disorder is a more specific term, but both may be deemed acceptable by a PBL tutor or facilitator, even in the context of the same problem scenario.

Acceptable solutions may also differ in the variety of reasoning paths that can be conceived, making medical PBL similar to the design problem discussed above. One medical expert interviewee reported that he formulated a PBL scenario of a patient with a known case of Lung Cancer with symptoms of Decreased Breath Sounds on One Side of the Chest. He expected the students to form the causal reasoning path in Figure 1(a).

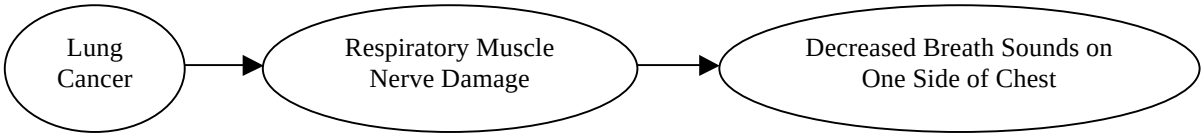


Fig. 1(a). Expert Causal Path.

However, the students came up with a different causal reasoning path, which was also correct and eventually deemed acceptable by the PBL tutor. The students formed the reasoning path shown in Figure 1(b).

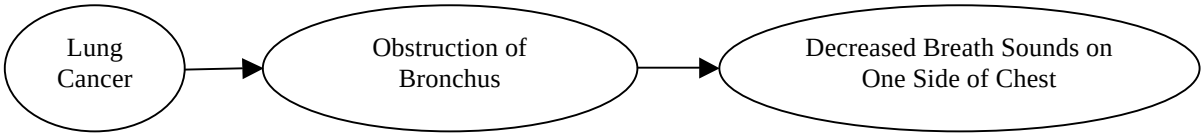


Fig. 1(b). Student Causal Path.

Our interviewees further reported that solutions designed or accepted by experts could also vary based on the knowledge of the targeted students. For example, students in their third year of medical studies would be expected to have greater knowledge of Pathology, whereas students in their first year may be asked to focus on disorders related to Anatomy.

The interviews also revealed that general difference of opinion in medical diagnosis can also occur due to the methods of clinical examination and the criteria of setting parameter thresholds. For example, a physician trained in the Manchester System of Classification (Tjandra, Clunie, Kaye & Smith, 2006) may diagnose a case of breast cancer as a third stage tumour, whereas another physician

trained in the system of the American Joint Committee on Cancer (AJCC) (American Joint Committee on Cancer, 2009), may diagnose the same case as a second stage tumour.

The differences in opinion highlighted above were also confirmed through the experimental feedback we received from medical experts. Different PBL tutors were found to disagree over the extent to which a causal link in a PBL hypothesis was acceptable. While one tutor found a causal link perfectly acceptable, another tutor was only inclined to accept it with reservation. Quite often a PBL tutor may accept varying solutions that differ in the choice of alternate terms or in the level of detail with which the solutions are presented. Our experiments also revealed that some experts were generally stricter than others in accepting plausible solutions.

SYSTEM PROTOTYPE

Our work is an extension of the COMET Collaborative Intelligent Tutoring System for Medical PBL (Suebnukarn & Haddawy, 2006; Kazi, Haddawy & Suebnukarn, 2007). In COMET, problem scenario solutions that form the system's domain models are authored by human domain experts. The student solution to a given problem scenario is compared to the stored expert solution and hints are generated if it deviates from the stored solution. However, this can discourage students from thinking creatively in analyzing the problem.

We have developed a novel approach to extend COMET to address this issue. Problem solutions collected from experts are combined with UMLS tables to form the domain model. The pedagogical module of the system consists of a hint generation mechanism that leverages off of the UMLS concept hierarchy and provides a probabilistic appraisal of causality between concepts to give students a measure of partial correctness of their hypotheses (Kazi et al., 2007). In this paper we focus on the issue of accommodating a greater variety of plausible solutions to a given problem scenario.

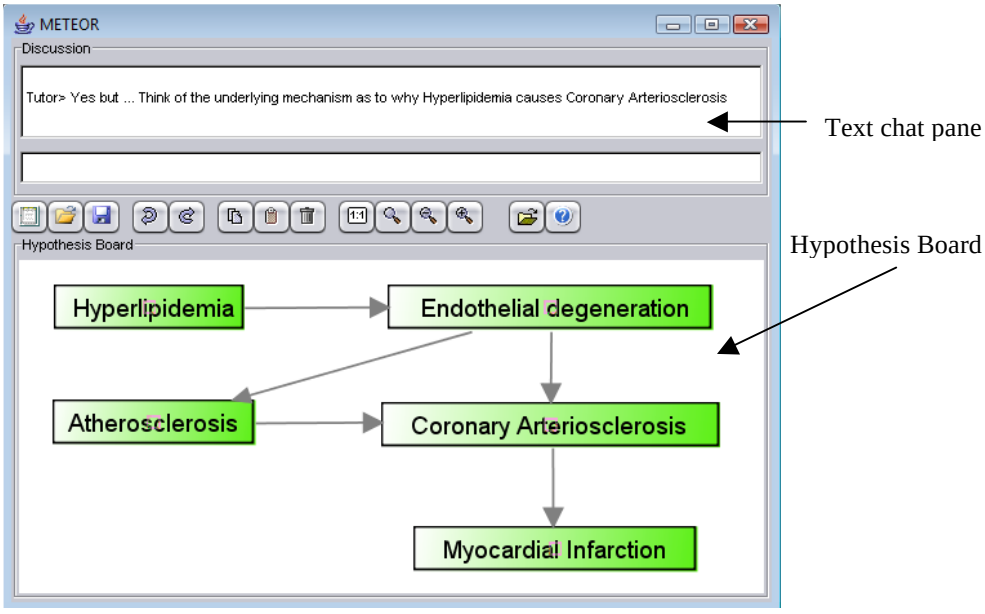


Fig. 2. System Prototype Interface.

The problem representation in our system is the same as that in COMET, a directed acyclic graph for forming the student hypothesis to a given problem. Similar to the COMET system interface, the student user is provided with a workspace as a hypothesis board to form his/her hypothesis, along with a text chat pane that returns hints from the system to guide the student in his/her clinical reasoning, as shown in Figure 2. The student chooses concepts from the UMLS Metathesaurus (U.S. National Library of Medicine, 2008) as hypothesis nodes, as shown in Figure 3. The concept search mechanism lists the various synonymous terms grouped together under a single concept to facilitate searches using synonyms. The student can draw edges between nodes through simple clicks of mouse buttons. The problem solving activity begins as the student is presented a problem scenario, such as:

> Mr. Heng-heng is a previously well, 48-year-old car dealer who is admitted with two hours history of severe crushing central chest pain. He has a poor controlled hypertension. He also gives a history of being tired recently and under stress. He has had indigestion for some weeks especially if walking after a heavy meal. His father died of a heart attack at age 55. He has smoked 20 cigarettes a day for 25 years. He is still in pain.

After studying the above problem description related to heart attack, the student hypothesizes that *endothelial degeneration* is a cause of *coronary arteriosclerosis*, which is shown to be a cause of *myocardial infarction*, as shown in Figure 2.

UMLS AS KNOWLEDGE SOURCE

The UMLS (U.S. National Library of Medicine, 2008) is a widely available medical knowledge source and is essentially a collation of various medical ontologies and terminologies (MeSH, SNOMED-CT, Gene Ontology, etc.). UMLS contains over 1 million medical concepts covering various medical domains and about 135 semantic types, where each medical concept is assigned at least one semantic type (U.S. National Library of Medicine, 2008). For example, the concept *vomiting* has the semantic type *finding*, whereas the concept *dehydration* has the semantic type *disease or syndrome*.

Our system makes use of the UMLS tables: *mrconso*, *mrsty*, *mrsab* and *mrrel*. The table *mrconso* contains the full list of medical concepts in the UMLS Metathesaurus. Synonymous concepts from different terminologies are grouped together under one unique concept and assigned a unique identifier. The table *mrsty* maps each unique concept to at least one semantic type (e.g., the semantic type of *diabetes* is *disease or syndrome*). The table *mrrel* lists the relationships between pairs of unique concepts, where relationships can be of various types such as *sibling*, *parent*, *child*, *broader*, *narrower*, *alike*, *synonymous*, *related*, and so on. An *alike* relationship exists between two concepts that are similar in meaning but have been assigned different unique identifiers and are thus separate concepts in UMLS. *Parent-Child* relationships are based on isa hierarchies. Concept *A* is the *parent* of concept *B*, if concept *A* is a general concept of which concept *B* is a specific type. For example, *vascular diseases* has a *parent* relationship with *atherosclerosis* and *atherosclerosis* has a *child* relationship with *vascular disease* as shown in Figure 4, while *coronary atherosclerosis* has an *alike* relationship with *coronary stricture*. The table *mrsab* maps a concept to its source terminologies such as Systematized Nomenclature for Medicine-Clinical Terms (SNOMED), Medical Subject Headings (MeSH), and so on.

Problem scenarios in medical PBL can span a diverse range of diseases and disorders within the broad medical domain, such as diabetes, pneumonia, heart attack, hepatitis, tuberculosis, head injury,

stroke, and so on. The range of medical concepts contained in the UMLS Metathesaurus is broad enough to cover the concepts referenced in the solutions to virtually any problem scenario for medical PBL. Thus, adding new problem scenarios to the tutoring system does not lead to additional encoding of new medical concepts to the domain ontology or system knowledge source, greatly reducing the burden of knowledge acquisition.

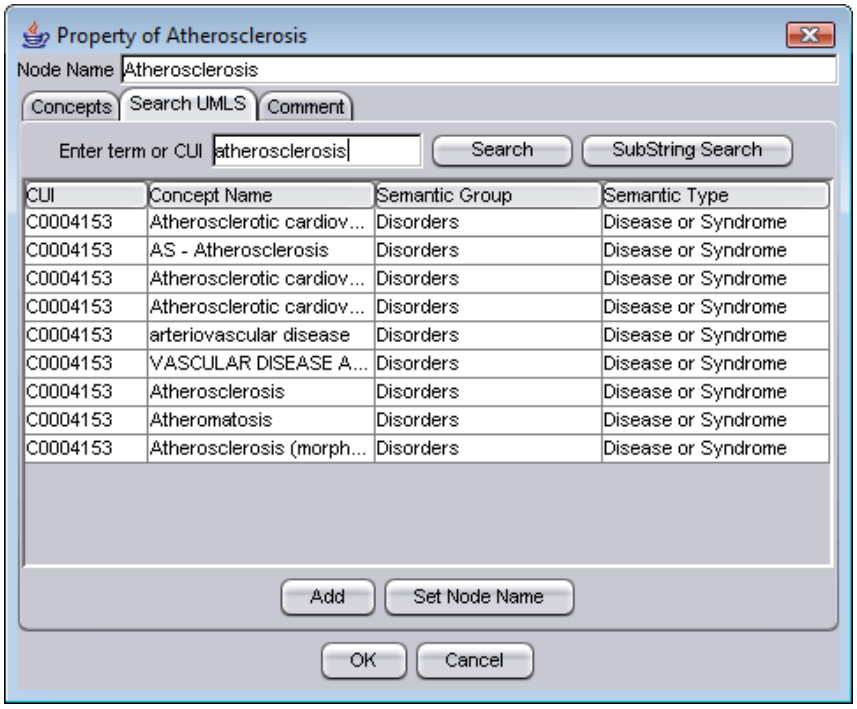


Fig. 3. System Interface for UMLS Concept Selection.

SYSTEM KNOWLEDGE BASE

The system knowledge base is formed by combining the relevant UMLS tables with solutions collected from human domain experts. The system architecture is shown in Figure 5. The system knowledge base is comprised of UMLS tables and an additional table that is henceforth referred to as the *expert knowledge base*. The *expert knowledge base* is encoded with the help of human domain experts, and it contains the causal relationship between various medical concepts, such as:

Hyperlipidemia → *Endothelial Degeneration*
Endothelial Degeneration → *Coronary Arteriosclerosis*
Endothelial Degeneration → *Atherosclerosis*
Coronary Arteriosclerosis → *Myocardial Infarction*

Student solutions that are considered acceptable by human experts are merged into the *expert knowledge base*. The *expert knowledge base* is formed through the collation of expert solutions to various problem scenarios, along with the student solutions that are certified by the experts to be

correct. The construction of an expert solution requires about 2-3 hours. Since each solution is in the form of a hypothesis graph, the collation of different solutions implies the incremental addition of the causal links in each solution to the *expert knowledge base*. The goal is to expand the *expert knowledge base* over time to include a greater number of expert approved solutions. This will facilitate the system in accepting a greater variety of plausible solutions.

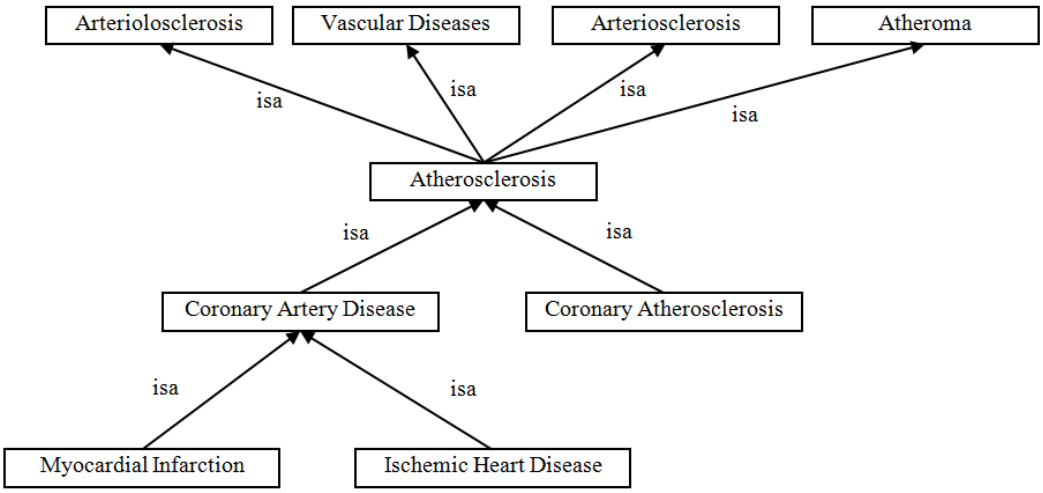


Fig. 4. Parent-Child Concept Hierarchy in UMLS.

SYSTEM REASONING & HYPOTHESIS EVALUATION

Each hypothesis causal link drawn by the student is evaluated by the system. The system refers to its knowledge base to check whether the link is acceptable. If the link is found to be acceptable, the system allows the directed edge (causal link) to be drawn; otherwise the system disallows the edge to be drawn and returns an appropriate hint to give feedback to the student. The acceptability of the student hypothesis link is evaluated by comparing it against the *expert knowledge base*. If there is a match, the link under evaluation is considered acceptable; however, if the link is not found in the *expert knowledge base*, the system makes use of a heuristic method to see if the link is acceptable.

This heuristic method makes use of the information structure within UMLS to make inferences. In expanding the set of plausible solutions, we make use of only the *alike* and *child-parent* relationships. Using the *alike* relationship, a concept in the PBL hypothesis can be replaced by similar meaning concepts, leading to a greater variety of plausible hypotheses. Using the *parent-child* relationship, a concept in the PBL hypothesis can be replaced by another concept that is broader and more general in meaning, leading to alternate hypotheses that differ in the level of detail or comprehensiveness. A hierarchy based on *parent-child* relationships between UMLS concepts is shown in Figure 4. While evaluating a causal link, we consider the *alike* and *parent-child* relationships of the source node and the destination node in the causal link. The source node in a causal link is the node from which the directed edge emanates, while the destination node is the node to which the directed edge leads. All concepts that are found to have *alike* or *parent* relationships with either the source or destination node are considered as alternatives to that node. For a causal link under

evaluation, let set *A* contain all concepts considered alternatives to the source node and let set *B* contain all concepts considered alternatives to the destination node. If the system finds a causal link in the *expert knowledge base* whose source node is a concept from set *A* and whose destination node is a concept from set *B*, then the link under evaluation will be considered acceptable.

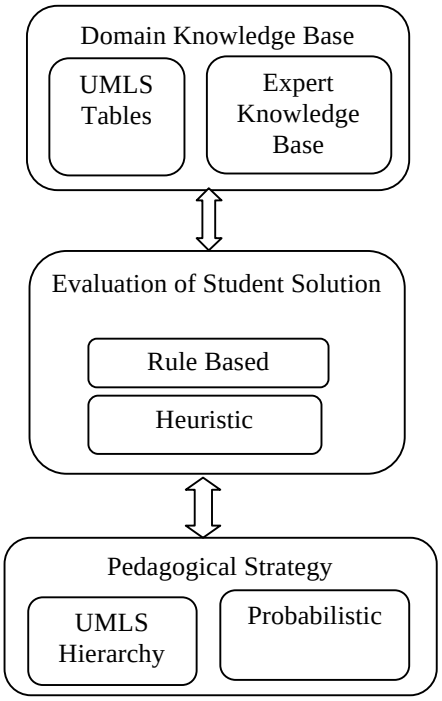


Fig. 5. System Architecture.

During our experimental evaluations, causal links having concepts from various medical terminologies were presented to medical experts. The experts were asked to indicate how acceptable they found the causal links presented to them by assigning a rating of acceptability. Our experiments revealed that links with concepts belonging to the terminologies of Medical Subject Headings (MSH), Systematized Nomenclature of Medicine-Clinical Terms (SNOMEDCT), National Cancer Institute (NCI) Thesaurus, and concepts created by NLM under the source abbreviation MTH, were found to be significantly more acceptable compared to other terminologies. Thus in order to reduce the number of erroneous links, the system only accepts causal links whose concepts belong to any of these sources.

Example

To illustrate the heuristic method of accepting causal links, consider the causal link from an expert solution for a problem scenario related to diabetes shown in Figure 6. The expert knowledge base contains a causal link leading from *hypoinsulinism* to *glucose metabolism disorder*. For this link, two lists are generated. The first list (L1) comprises concepts that have *alike* or direct *parent* relationships with *hypoinsulinism*, so that L1 = {*hypoinsulinism*, *diseases of endocrine pancreas*}. The second list (L2) comprises concepts that have *alike* or direct *parent* relationships with *glucose metabolism*

disorder, so that L2 = {*glucose metabolism disorders*, *metabolic diseases*, *disorder of carbohydrate metabolism*}. Here, *metabolic diseases* and *disorder of carbohydrate metabolism* are both direct parents of *glucose metabolism disorders*. Now the system accepts all hypothesis links that are formed through L1 → L2, such as:

hypoinsulinism → *glucose metabolism disorder*
hypoinsulinism → *metabolic diseases*
hypoinsulinism → *disorder of carbohydrate metabolism*
diseases of endocrine pancreas → *glucose metabolism disorder*
diseases of endocrine pancreas → *metabolic diseases*
diseases of endocrine pancreas → *disorder of carbohydrate metabolism*

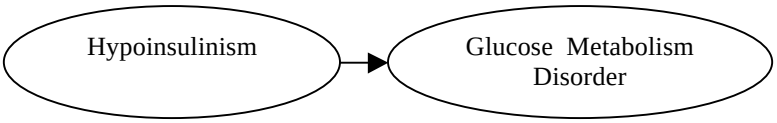


Fig. 6. Expert Causal Link.

Thus, using this heuristic method of expanding the solution set, the system is able to expand the space of plausible solutions.

HYPOTHESIS GRAPH EVALUATION

The tutoring system starts by evaluating each causal link in the hypothesis separately. After the causal links have been evaluated and corrected, the system evaluates the graph as a whole, following a student request made through the click of a button on the system interface. The system begins with the hypothesis nodes that represent the patient symptoms, and then reasons backwards to make sure that a valid hypothesis has been constructed. A hypothesis is considered valid if the patient symptoms from the problem description and their respective causing factors have been adequately represented. A valid hypothesis should contain a chain of reasoning, where nodes representing enabling conditions lead to other nodes in succession that eventually lead to the symptoms.

The system begins by checking for symptom nodes from the problem scenario. If a symptom is not found in the student hypothesis, the system responds with an appropriate hint. After ensuring that all symptoms are duly represented in the student hypothesis, the system begins to reason backwards for each of the symptoms. This is achieved by checking that each node that does not represent an enabling condition is shown to be a cause of other nodes. The system performs a recursive search through the chain of nodes to ensure that the root nodes in the hypothesis graph are nodes that represent enabling conditions. For example, based on the problem scenario related to heart attack described above, the path of reasoning shown in Figure 7 will be validated as follows.

The system begins with *chest pain* as it is a declared symptom in the problem scenario. The system then verifies that *chest pain* is caused by *unstable angina*, *unstable angina* is caused by *thrombosis*, *thrombosis* is caused by *coronary arteriosclerosis*, *coronary arteriosclerosis* is caused by *endothelial degeneration* and *endothelial degeneration* is caused by *hyperlipidemia*. Since a known

cause of the root node of hyperlipidemia is not fed to the system, the system assumes it as an enabling condition. Thus the system validates this reasoning path as being acceptable.

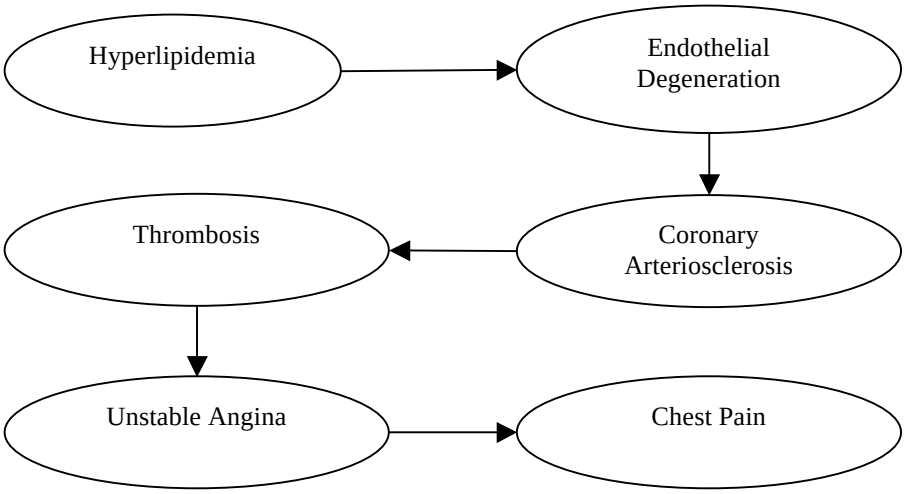


Fig. 7. Example Hypothesis Graph.

RESULTS

We conducted evaluations of the acceptability of causal links generated by our heuristic method. In order to exhaustively explore the space of links accepted by the system, we generated all causal links that would potentially be accepted by the system, based on a given sample of expert links. We randomly selected 14 causal links from expert solutions to three problem scenarios in heart attack, diabetes and pneumonia. For each node in the causal link we generated a list of concepts that were found to have *alike* and *parent* relationships with the nodes in the causal link. Thus, we generated a pair of lists of concepts for each causal link. We then formed links between each pair of concepts found in the respective lists spanning the full permutation of the pair of lists.

From the initial 14 links, the system generated a total of 228 links based on the *alike* and *parent* relationships. These system-generated causal links were then presented to 10 medical experts from Thammasat University Medical School, who had at least 5 years of experience in conducting PBL sessions. The medical experts were informed about each problem scenario to which the causal links were related. They were asked to rate the acceptability of each link on a scale of 1-5, where 1 meant unacceptable, 2 meant not quite acceptable, 3 meant not sure, 4 meant close to acceptable and 5 meant acceptable. The ratings were so chosen to accommodate the difference of opinion often found among PBL tutors in the medical domain, as discussed earlier. Figure 8 shows part of the expert evaluation form based on the links generated from the causal link shown in Figure 6.

Diabetes Case

Causal Links	Acceptability
Hypoinsulinism-->Glucose Metabolism Disorders	5 4 3 2 1
Hypoinsulinism-->Metabolic Diseases	5 4 3 2 1
Hypoinsulinism-->Disorder of carbohydrate metabolism	5 4 3 2 1
Disorder of endocrine pancreas-->Glucose Metabolism Disorders	5 4 3 2 1
Disorder of endocrine pancreas-->Metabolic Diseases	5 4 3 2 1
Disorder of endocrine pancreas-->Disorder of carbohydrate metabolism	5 4 3 2 1

Fig. 8. Part of Expert Evaluation Form.

In order to have a set of samples that are representative of student knowledge and reasoning, these 228 links were short listed by medical experts to comprise only links that could conceivably be formed by medical students according to expert judgment. This resulted in 213 links, which were further short listed to consist of only those links whose concepts belonged to any of the above mentioned four terminologies: MSH, SNOMEDCT, MTH and NCI.

Thus a total of 111 system generated causal links were considered for evaluation. Based on the ratings 1-5 assigned by the medical experts, we computed the average score for each causal link. Of these 111 links, 43 links were based on the heart attack case, 25 links were based on the diabetes case and 43 links were based on the pneumonia case. The overall mean of the scores came out to be 4.11 with a standard deviation mean of 0.69.

Table 1
System Acceptability of Causal Links

Acceptability Threshold	Acceptable Links	Unacceptable Links	Accuracy
3.0	106	5	95.49 %
3.5	91	20	81.98 %
4.0	76	35	68.46 %

To get a measure of acceptability we collapsed the rating scale to divide the 111 links into acceptable or unacceptable. All links that had an average score equal to or above a particular threshold were considered acceptable, while the rest were considered unacceptable. Table 1 shows the system accuracy for varying thresholds of acceptability.

For the total number of 111 samples, the experts were found to agree with each other with a good degree of statistical agreement (Pearson Correlation Coefficient = 0.48, $p < 0.05$). Figure 9 shows the mean value of scores assigned by individual experts. As can be seen in Figure 9, some experts were stricter than others in scoring inferred links, with mean scores ranging from 3.6 to 4.7. Figure 10 shows the correlation values of individual experts with the rest of the experts. Since the rating categories presented to experts were ordinal in nature, we chose to compute the inter-expert correlation.

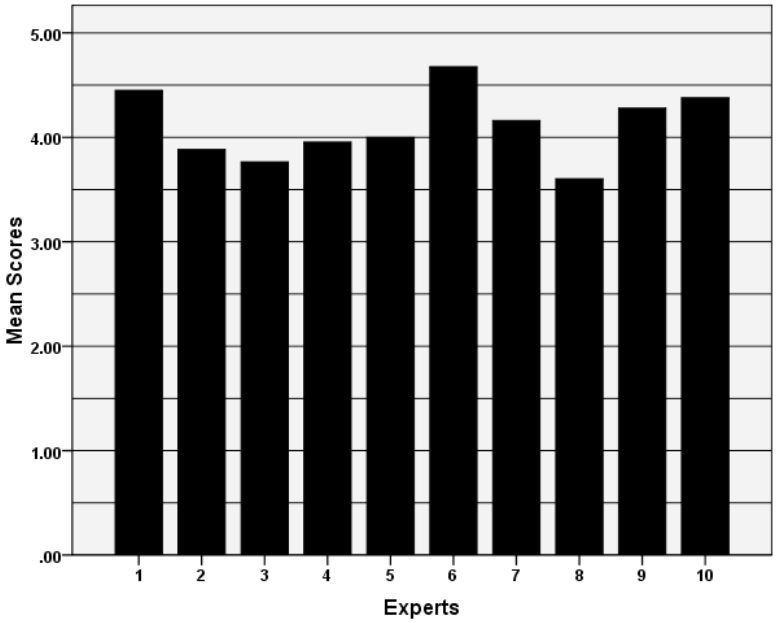


Fig. 9. Mean Values of Scores by Individual Experts.

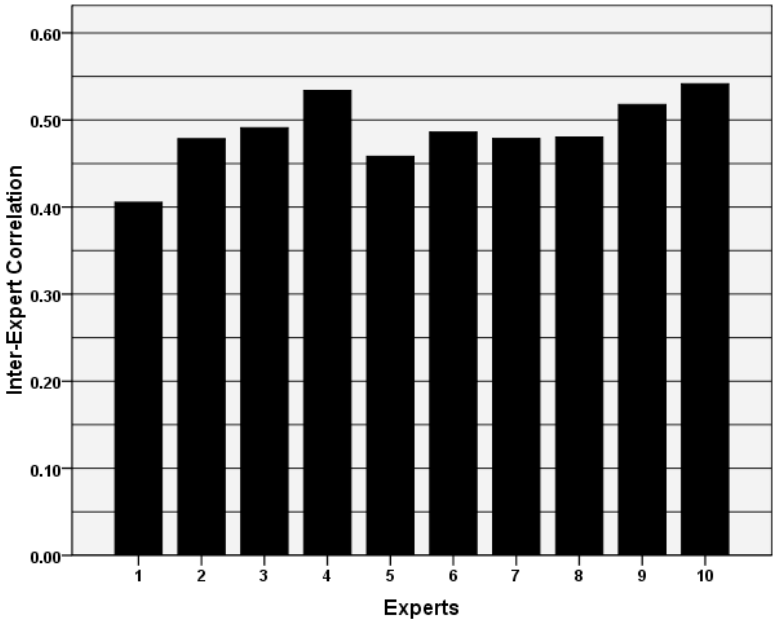
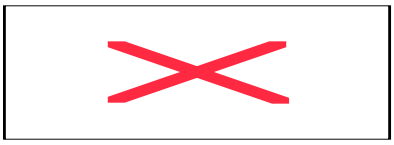


Fig. 10. Correlation Values of Individual Experts with Remaining Experts.

Precision & Recall

In the information retrieval task, a document that is relevant but does not contain the specific query term may not be retrieved. This mismatch can be addressed by using a thesaurus to expand the query so that it contains a greater number of terms that are similar in meaning to the original query term. This increases the likelihood of matching a relevant document that may not contain the original query term, but instead contains alternate terms. Performance results of information retrieval are often reported in terms of precision and recall. Query expansion using a thesaurus often results in increased recall.

We infer concepts from the UMLS using *alike* and *parent-child* relationships to expand the original base of an approved causal link, thereby admitting a greater number of plausible links. Thus, we liken our task of expanding the plausible solution space using UMLS to the information retrieval task of using a thesaurus for query expansion. Therefore to describe the system performance, we also report the precision and recall, which in our context can be defined as:



$$\text{Precision} = \frac{\text{Acceptable} \cap \text{Inferred}}{\text{Inferred}}$$

In order to compute the total number of acceptable links that serve as alternatives for the original sample of 14 links, we conducted a separate expert evaluation. Four medical experts from Isra University were asked to identify possible alternatives for each of the 14 causal links. The experts identified only a handful of alternatives initially, but were willing to accept a wider variety of plausible alternatives when presented with them. Thus, to get a rough estimate of the total number of possible alternatives, the experts were finally asked to provide only an estimate of the total number of possible alternatives for each of the 14 causal links. The average estimate of the possible alternatives for the combined 14 links came out to be 245. Thus the experts estimated that there could be roughly 245 links that would be considered acceptable alternatives of the original sample of 14 links. Based on the estimated figure, the precision and recall values of our system for the acceptability threshold of 4.0 are:

$$\Rightarrow \text{Recall with expanded solutions} = \frac{76}{245} = 0.3102$$

$$\Rightarrow \text{Precision with expanded solutions} = \frac{76}{111} = 0.6846$$

A tutoring system that only accepts explicitly encoded expert solutions will have the following precision and recall values:

$$\Rightarrow \text{Recall with expert solution} = \frac{14}{245} = 0.0571$$

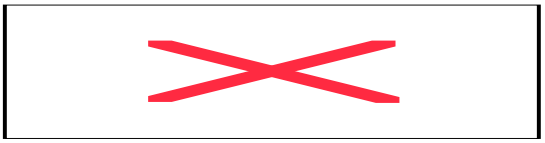
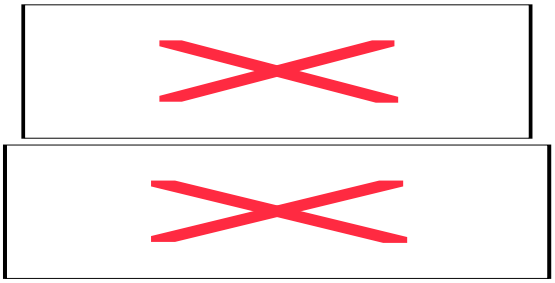


Table 2 shows a comparison of precision and recall values for our system against those of a tutoring system without the expanded solution space. The precision and recall values shown in this table are based on the acceptability threshold of 4.0.

Table 2
Precision & Recall: System with Expanded Solutions vs. System without Expanded Solutions

	Precision	**Recall**
With Expanded Solutions	0.6846	0.3102
Without Expanded Solutions	1.0	0.0571

Regardless of the number of causal links estimated by experts to be alternatives for the original sample of 14 links, the comparison of our system with expanded solutions against a tutoring system without expanded solution space vis-à-vis precision and recall, still holds. The recall of our system with expanded solutions still comes out to be more than five times the recall of a tutoring system that only accepts expert solutions. Let X be the total number of alternative causal links, the recall will be:



In order to compare the performance of our system with human experts, we computed the precision values for each of the ten experts. The rating assigned by an individual expert for a link was compared against the average expert score for that particular link. We took the acceptability threshold of 4.0 as the gold standard: all links having an average score of 4 or above were considered acceptable, while the rest were considered unacceptable. An expert rating of 4 or 5 implied that the link was accepted by the expert, otherwise the link was considered to have been deemed unacceptable by the expert. We computed expert precision as:

$$\text{Expert Precision} = \frac{\text{Acceptable by Gold Standard} \cap \text{Expert Rating of 4 or 5}}{\text{Expert Rating of 4 or 5}}$$

Thus, the expert precision indicates the proportion of links that were deemed acceptable based on the gold standard and also rated acceptable by the expert, out of the total number of links rated acceptable by the expert. Figure 11 shows the resulting precision values of individual experts for the acceptability threshold of 4.0. The precision of experts varies from a minimum of 70% to a maximum of 91%, with an average precision of about 82%.

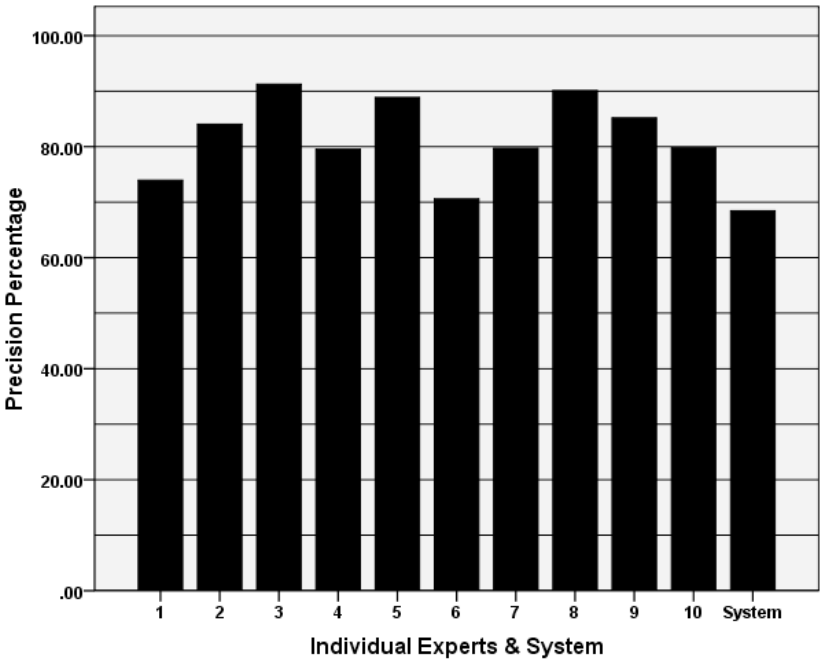


Fig. 11. Precision scores of individual experts & our system.

Receiver Operating Characteristic (ROC) Analysis

ROC analysis is often used to report classifier performance along the dimensions of false positive rate (1-specificity) and true positive rate (sensitivity) (Fawcett, 2006), which are defined as:

$$\text{True Positive Rate} = \frac{\text{Positive Instances Classified as Positive}}{\text{Total Positive Instances}}$$

$$\text{False Positive Rate} = \frac{\text{Negative Instances Classified as Positive}}{\text{Total Negative Instances}}$$

In order to compare the classification ability of our system with a tutoring system that only accepts expert solutions, we present an ROC analysis of the discrete classifiers, by using the fall out

measure (false positive rate) and recall (true positive rate). In our context, the fallout can be defined as:

$$\text{Fallout} = \frac{\text{Unacceptable} \cap \text{Inferred}}{\text{Unacceptable}}$$

To estimate the total number of causal links that would be deemed unacceptable, we simply subtract the total estimated number of acceptable links from the rough estimate of the total number of links that could possibly be formed for a medical PBL hypothesis using concepts from UMLS. The UMLS concepts that help form a medical PBL hypothesis generally belong to the semantic types: *Finding*, *Sign or Symptom*, *Diseases or Syndrome*, *Pathologic Function*, *Molecular Function*, *Organ or Tissue Function*, or *Organism Function*. The number of UMLS concepts belonging to any of these semantic types is roughly 10^5. Thus the total number of causal links possibly formed through these concepts is roughly 10^5 x 10^5. We compute the system fallout for the acceptability threshold of 4.0:

$$\Rightarrow \text{Fallout with expanded solutions} = \frac{35}{(10^5 \times 10^5) - 245} = \frac{35}{10^{10} - 245} = 3.5 \times 10^{-9}$$

The fallout of a tutoring system that only accepts expert solutions is approximately zero, since the solutions have been certified by experts to be correct. Likewise the recall of such a tutoring system based on the sample of 14 causal links will be:

$$\text{Recall with expert solution} = \frac{14}{245} = 0.0571$$

Table 3 shows the discrete classifiers of our system with expanded solutions and a tutoring system without expanded solutions, in the ROC space.

Table 3
ROC Analysis

	FP Rate (1 – Specificity)	**TP Rate (Sensitivity)**	**Geometric Mean** $\left(\sqrt{Specificity \bullet Sensitivity}\right)$
System With Expanded Solutions	3.5 x 10^{-9}	0.3102	0.56
System Without Expanded Solutions	0	0.0571	0.23

Precision with Student Generated Links

In order to test the system accuracy with causal links generated by students, we collected a sample of 211 links that were drawn by six 3rd year medical students during a two hour trial evaluation of the system. The six students were divided into three groups who worked on the problem scenarios of diabetes, heart attack and pneumonia. From these 211 links, 61 were accepted by the system, out of

which 12 links were accepted by the system through the inference mechanism. To test the precision of the system's inference mechanism, these 12 links were presented to 5 medical experts at Isra University for an acceptability rating. The average score came out to be 4.58 with a standard deviation mean of 0.47. The precision of the individual experts and the system based on these 12 links is shown in Figure 12.

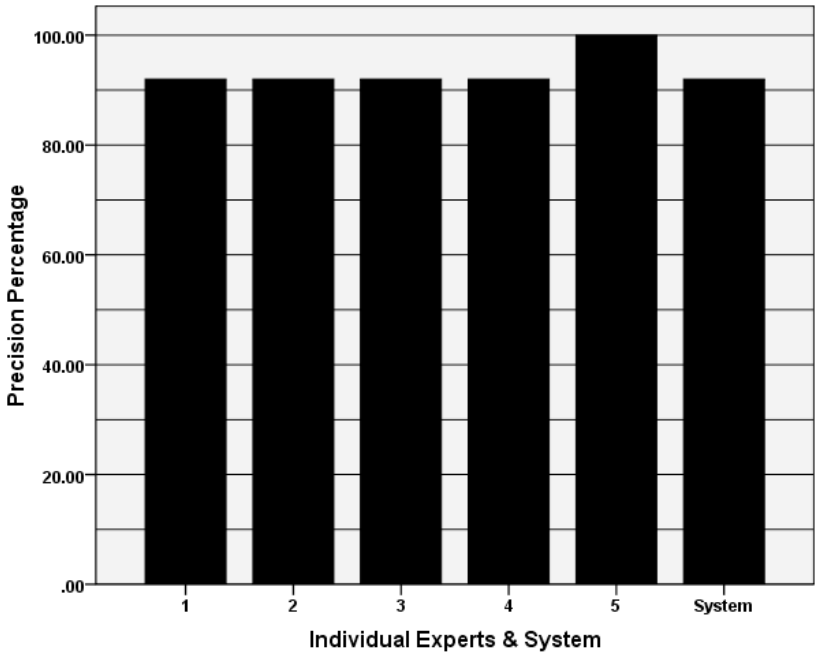


Fig. 12. Precision of individual experts & system.

DISCUSSION

While the system precision with student generated links is much higher, we believe the measures obtained through system generated links are more reflective of the system's capability. This is because system generated links give an exhaustive coverage of links that can possibly be accepted by the system through inference. On the other hand, student generated links are highly dependent on the profile of the students. For instance, the students in our trial were at the end of their 3rd year studies and had strong subject knowledge. Their causal links that were accepted by the system through inference were expected to be sound in correctness and indeed received strong acceptance from human experts, too. However, that may not be the case with students with lesser subject knowledge, who may draw incorrect links that may still be accepted by the system through inference. Thus, we focus our remaining discussion on the measures obtained through the testing of system generated links.

The average score of 4.11 out of 5.0 and good system accuracy for varying thresholds of acceptability indicate fair expert acceptance of the system generated solutions. While the precision obtained using the expanded solution is less than that using the explicitly encoded solution, this is compensated for by the significant increase in recall. By losing a bit in precision, we are gaining by

providing students more scope and room for creativity in constructing their solutions. In interpreting the reduced precision of the expanded solution, it is also worth noting that the inter-expert correlation of 0.48 means that there is a fair amount of disagreement among experts concerning the acceptability of solutions. For instance, 10 out of the 35 links found unacceptable received an acceptability score of 5 by three or more experts.

The inferred links receiving the least scores were further examined to investigate the reasons behind the reduced precision. Expert insight revealed that the inferred concepts were in most cases over generalized or non-representative of the original concept, meaning that the parent concept in the inferred link was semantically quite distant from its child concept in the expert approved link. For example, the expert approved link *Dehydration* → *Thirsty* led to the inferred links *Fluid & Electrolyte Manifestation* → *Thirsty* and *Hydration Status* → *Thirsty*. Here, *Fluid & Electrolyte Manifestation* is a parent of *Dehydration*, but is not truly representative of dehydration because the term manifestation is too broad. Similarly the term *Hydration Status* is too vague as it could possibly imply dehydration or overhydration. We believe the reason behind this over-generalization of parent concepts lies in the way parent-child relationships are constructed in the source vocabularies in UMLS. If the parent-child hierarchy were more tightly coupled such that a direct parent was closer in meaning to its child, then such an ontology would likely facilitate the inference of causal links that were more plausible.

The geometric mean of specificity and sensitivity shown in the ROC analysis in Table 3 indicates that the classifier of our system supersedes that of a tutoring system without an expanded solution space. This is because the true positive rate of our system is significantly higher, while the false positive rate is almost the same for both systems (Fawcett, 2006).

As can be observed from the mean scores and correlation values in Figures 9 and 10, the experts were not in perfect agreement with one another over the acceptability of the causal links, which is quite characteristic of an ill-defined domain such as medicine. Furthermore, the expert evaluations also revealed that in some cases, links inferred by the system scored higher than the corresponding original links created by human experts. For example, two system inferred links: *heredity* → *vascular diseases* and *heredity* → *arteriosclerosis* received mean scores of 4.5 and 4.2 respectively, whereas the original expert link *heredity* → *atherosclerosis* received a mean score of 3.3. Thus, a causal link created by some experts was not considered acceptable or even close to acceptable by other experts, whereas a set of links inferred by the system off of the same expert link was considered by experts to be very close to acceptable. This evidence further points to the inherent variation and subjectivity involved in the evaluation of medical PBL hypothesis links by human experts. This also shows that for the formation of a hypothesis, experts are not always able to select the best term or concept and that the use of the UMLS Metathesaurus can be quite helpful in this regard. The utility of the UMLS Metathesaurus is equally applicable or perhaps even more so in the case of students forming their hypotheses. This inherent subjectivity in the evaluation of medical PBL hypotheses and the prospects of alternate terms reinforces the need to have a tutoring system that allows students to work creatively and adopt novel solutions, and that also evaluates student solutions in a broad context.

CONCLUSIONS AND FUTURE WORK

In this paper we have addressed the issue of classifying problem solutions as correct or incorrect in tutoring systems in the context of an ill-defined domain such as medical problem-based learning. We have described how to reduce the burden of knowledge acquisition in an intelligent tutoring system by

deploying an existing domain knowledge source. We have described how a broad and easily available medical knowledge source such as UMLS can be deployed as the domain ontology for a tutoring system for medical PBL. We have presented a strategy of making the tutoring system more robust by broadening the scope of solutions that are accepted by the tutoring system as being correct. Terms or concepts in expert certified medical PBL solutions are replaced by synonymous or broader meaning terms in UMLS to expand the space of plausible solutions. We have illustrated through examples how the tutoring system can be made more robust in evaluating various plausible student solutions. Our approach is innovative in exploiting the UMLS information structure to expand the space of plausible solutions, while at the same time widening the knowledge acquisition bottleneck.

Although the design and implementation of our tutoring system is focused on problem-based learning in the medical domain, the inference mechanism described in this paper can also be applied to other ill-defined domains where the problem representation is in the form of causal graphs. Graph nodes representing domain concepts, can be selected from the domain ontology, where graph edges describe the cause and effect relationship between two domain concepts. Based on concept relationships defined in the domain ontology or thesaurus, a given concept node in a causal graph may be replaceable by other similar concepts, thus leading to a greater variety of plausible causal graph solutions to a given problem. The use of a domain ontology or thesaurus also assists the users whether expert or student, in selecting alternate terms or concepts that are more suitable for the formation of a particular hypothesis.

We would also like to evaluate the system performance for problem scenarios other than the three cases of diabetes, heart attack and pneumonia. We intend to evaluate the effectiveness of the system's tutoring hints vis-à-vis the two major components of hint generation in our system: the measure of partial correctness and the leveraging off of the UMLS concept hierarchy. Finally, we intend to conduct evaluations of learning outcomes by assessing the clinical reasoning gains acquired by student users as a result of using this medical tutoring system. As an extension to the existing work, we would like to allow students to be more creative in forming novel solutions and have the system evaluate their plausibility, which is a significantly more challenging problem.

ACKNOWLEDGEMENTS

We would like to thank the medical experts at Thammasat University and Isra University for donating their time and effort during the evaluations. We extend a special note of appreciation to Dr. Dev Anand and Dr. Afroz S. Kazi for providing technical feedback related to ill-definedness of medical PBL and the analysis of causal links that led to reduced precision. We also thank the medical students of Isra University who donated their time during the system trial evaluation. Finally we are thankful to the reviewers whose thorough comments helped to improve the manuscript.

REFERENCES

Achour, S. L., Dojat, M., Rieux, C., Bierling, P., & Lepage, E. (2001). A UMLS-based knowledge acquisition tool for rule-based clinical decision support systems development. *Journal of the American Medical Informatics Association*, 8 (4), 351-360.

Albanese, M. A. (2004). Treading tactfully on tutor turf: Does PBL tutor content expertise make a difference? *Medical Education, 2004,* 38, 916-920.

Aleven, V., Pinkwart, N., Lynch, C. F., & Ashley, K. D. (2006). Supporting self-explanation of argument transcripts: Specific v. generic prompts. In V. Aleven, K. Ashley, C. Lynch, & N. Pinkwart (Eds.) *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems* (pp. 47-55). Berlin: Springer-Verlag.

American Joint Committee on Cancer (2009). *American Joint Committee on Cancer*. Retrieved from http://www.cancerstaging.org/.

Anderson, J. R., Corbett, A., Koedinger, K., & Pelletier, R. (1996). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences,* 4 (2), 167-207.

Burgun, A., & Bodenreider, O. (2001). Methods for exploring the semantics of the relationships between co-occurring UMLS concepts. *MedInfo,* 10(Pt 1), 171-175.

Crowley, R., & Medvedeva, O. (2006). An intelligent tutoring system for visual classification problem solving. *Artificial Intelligence in Medicine,* 36 (1), 85-117.

Crowley, R. S., Tseytlin, E., & Jukic, D. (2005). ReportTutor - an intelligent tutoring system that uses a natural language interface. *Proceedings of the American Medical Informatics Association Symposium, 2005* (pp. 171-175).

Das, M., Mpofu, D. F. S., Hasan, M. Y., & Stewart, T. S. (2002). Student perceptions of tutor skills in problem-based learning tutorials. *Medical Education* 36, 272-278.

Day, M. Y., Lu, C., Yang, J. D., Chiou, G., Ong, C. S., & Hsu, W. (2005). Designing an ontology-based intelligent tutoring agent with instant messaging. In P. Goodyear (Ed.) *Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies, ICALT 2005* (pp. 318-320). Washington, DC: IEEE Computer Society.

Dragon, T., & Woolf, B. P. (2006). Guidance and collaboration strategies in ill-defined domains. In V. Aleven, K. Ashley, C. Lynch & N. Pinkwart (Eds.) *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems,*(pp. 65-73). Berlin: Springer-Verlag.

Easterday, M. W., Aleven, V., & Scheines, R. (2007). The logic of Babel: Causal reasoning from conflicting sources. In V. Aleven, K. Ashley, C. Lynch & N. Pinkwart (Eds.) *Proceedings of the Workshop on AIED Applications in Ill-Defined Domains at the 13th International Conference on Artificial Intelligence in Education* (pp. 31-40).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters,* 26, 861-874.

Feltovich, P. J., & Barrows, H. S. (1984). Issues of generality in medical problem solving. In H. G. Schmidt & M. L. De Volder (Eds.) *Tutorials in Problem-BasedLlearning: A New Direction in Teaching the Health Professions.* The Netherlands: Van Gorcum.

Gauthier, G., Lajoie, S. P., & Richard, S. (2007). Mapping and validating case specific cognitive models. In V. Aleven, K. Ashley, C. Lynch & N. Pinkwart (Eds.) *Proceedings of the Workshop on AIED Applications in Ill-Defined Domains at the 13th International Conference on Artificial Intelligence in Education, 2007.*

Hay, P. J., & Katsikitis, M. (2001). The 'expert' in problem based and case-based learning: necessary or not? *Medical Education, 2001,* 35, 22-26.

Hersh, W. R., Price, S., & Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS Metathesaurus. In J. M. Overhage (Ed.) *Proceedings of the Annual American Medical Informatics Association Symposium, 2000* (pp. 344-348). Philadelphia: Hanley and Benfus.

Jonassen, D. H. (1997). Instructional design models for well-structured and iII-structured problem-solving learning outcomes. *Educational Technology Research and Development,* 45, 65-94.

Kazi, H., Haddawy, P., & Suebnukarn, S. (2007). Enriching solution space for robustness in an intelligent tutoring system. In T. Hirashima, H.U. Hoppe & S. Shwu-Ching Young (Eds.) *Proceedings of 15th International Conference on Computers in Education* (pp. 547-550). Amsterdam, The Netherlands: IOS Press.

Kazi, H., Haddawy, P., & Suebnukarn, S. (2007). Towards human-like robustness in an intelligent tutoring system. In R. L. Lewis, T. A. Polk & J.E. Laird (Eds.) *Proceedings of Eighth International Conference on Cognitive Modeling* (pp. 247-252). Oxford, UK: Taylor & Francis/Psychology Press.

Kumar, A. (2002). Model-based reasoning for domain modeling in a web-based intelligent tutoring system to help students to learn to debug C++ programs. In S.A. Cerri, G. Gouarderes & F. Paraguacu (Eds.) *Proceedings of the 6th International Conference on Intelligent Tutoring Systems, 2002* (pp. 792-801). Berlin: Springer-Verlag.

Le, N. T., & Menzel, W. (2007). Using constraint based modelling to describe the solution space of ill-defined problems in logic programming. In H. Leung, F. Li, R. Lau & Q. Li (Eds.) *Proceedings of the 6th International Conference on Web-based Learning* (pp. 367-379). Berlin, : Springer.

Lee, C. H., Seu, J. H., & Evens, M. W. (2002). Building an ontology for CIRCSIM tutor. In *Proceedings of the 13th Midwest Artificial Intelligence and Cognitive Science Society Conference, MAICS 2002* (pp. 161-168).

Liu, Z., & Chu, W. (2007). Knowledge-based query expansion to support scenario-specific retrieval of medical free text. *Information Retrieval,* 10(2), 173-202.

Lynch, C., F., Ashley, K., D., Aleven, V., & Pinkwart, N. (2006). Defining ill-defined domains: A literature survey. In V. Aleven, K. Ashley, C. Lynch & N. Pinkwart (Eds.) *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems* (pp. 1-10). Berlin: Springer-Verlag.

Martens, A., Bernauer, J., Illmann, T., & Seitz, A. (2001). Docs 'n Drugs – The virtual polyclinic: An intelligent tutoring system for web-based and case-oriented training in medicine. In S. Bakken (Ed.) *Proceedings of American Medical Informatics Association Symposium, 2001* (pp. 433-437). Philadelphia: Hanley and Belfus.

Mendonca, E. A., & Cimino, J. J. (2000). Automated knowledge extraction from MEDLINE citations. In J. M. Overhage (Ed.) *Proceedings of American Medical Informatics Association Symposium, 2000* (pp. 575-579). Philadelphia: Hanley and Belfus.

Mills, B., Evens, M., & Freedman, R. (2004). Implementing directed lines of reasoning in an intelligent tutoring system using the atlas planning environment. *Proceedings of the International Conference on Information Technology: Coding and Computing* (pp. 729-733).

Mitrovic, A. (1998). Experiences in implementing constraint-based modelling in SQL-Tutor. In B. P. Goettle, H. M. Halff, C. L. Refield & V. J. Shute, (Eds,) *Proceedings of the 4th International Conference on Intelligent Tutoring Systems, 1998* (pp. 414-423). London, UK: Springer-Verlag.

Murray, T. (1999). Authoring intelligent tutoring systems: An analysis of the state of the art. *International Journal of Artificial Intelligence in Education,* 10, 98-129.

National Cancer Institute. (2009). *NCI Metathesaurus Browser Home Page*. Retrieved from http://ncimeta.nci.nih.gov/MetaServlet/

Oguejiofor, E., Kicinger, R., Popovici, E., Archiszewski, T., & Jong, K. D. (2004). Intelligent tutoring systems: An ontology based approach. *International Journal of Information Technology in Architecture, Engineering and Construction,* 2 (2), 115-128.

Pople, H. E., Jr. (1982). Heuristic methods for imposing structure on ill-structured problems: The structuring of medical diagnostics. In P. Szolovits (Ed.) *Artificial Intelligence in Medicine* (pp. 119-190). *Boulder, CO:* Westview Press.

Remolina, E., Ramachandran, S., Fu, D., Stottler, R., & Howse, W. (2004). Intelligent simulation-based tutor for flight training. In *Proceedings of the Industry/Interservice, Training, Simulation & Education Conference, 2004.*

Rubin, S. H., Rush Jr., R. J., Smith, M. H., Murthy, S. N. J., & Trajkovic, Lj. (2002). A soft expert system for the creative exploration of first principles of crystal-laser design. *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, SMC 2002, Hammamet, Tunisia, Oct. 2002.*

Srinivasan, S., Rindflesch, T. C., Hole, W. T., Aronson, A. R., & Mork, J. G. (2002). Finding UMLS Metathesaurus concepts in MEDLINE. In I. S. Kohane (Ed.) *Proceedings of American Medical Informatics Association Symposium, 2002* (pp. 727-731). Philadelphia: Hanley and Belfus.

Suebnukarn, S., & Haddawy, P. (2006). Modeling individual and collaborative problem-solving in medical problem-based learning. *User Modeling and User Adapted Interaction,* 16 (3), 211-248.

Suebnukarn, S., & Haddawy, P. (2007). COMET: A collaborative tutoring system for medical problem-based learning. *IEEE Intelligent Systems*, 22(4), 70-77.

Suraweera, P., Mitrovic, A., & Martin, B. A. (2005). Knowledge acquisition system for constraint based intelligent tutoring systems. In C. K. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.) *Proceedings of the International Conference on Artificial Intelligence in Education, 2005* (pp. 638-645). Amsterdam, The Netherlands: IOS Press.

Tjandra, J. T., Clunie, G. J. A., Kaye, A. H., & Smith, J. A. (2006). *Textbook of Surgery* (3rd Ed.). Oxford: Blackwell Publishing, 2006.

U.S. National Library of Medicine. (2008). *About the UMLS Resources*. Retrieved from http://www.nlm.nih.gov/research/umls/about_umls.html