Journal of Biomedical Informatics 45 (2012) 557–565

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Employing UMLS for generating hints in a tutoring system for medical problem-based learning

Hameedullah Kazi [a,*], Peter Haddawy [b], Siriwan Suebnukarn [c]

[a] Department of Electrical Engineering & Computer Science, Isra University, Pakistan
[b] United Nations University International Institute for Software Technology, Macao
[c] School of Dentistry, Thammasat University, Thailand

ARTICLE INFO

Article history:
Received 15 July 2011
Accepted 28 February 2012
Available online 13 March 2012

Keywords:
Ontology
Hint generation
Intelligent tutoring systems
Medical PBL
UMLS
Knowledge acquisition bottleneck

ABSTRACT

While problem-based learning has become widely popular for imparting clinical reasoning skills, the dynamics of medical PBL require close attention to a small group of students, placing a burden on medical faculty, whose time is over taxed. Intelligent tutoring systems (ITSs) offer an attractive means to increase the amount of facilitated PBL training the students receive. But typical intelligent tutoring system architectures make use of a domain model that provides a limited set of approved solutions to problems presented to students. Student solutions that do not match the approved ones, but are otherwise partially correct, receive little acknowledgement as feedback, stifling broader reasoning. Allowing students to creatively explore the space of possible solutions is exactly one of the attractive features of PBL. This paper provides an alternative to the traditional ITS architecture by using a hint generation strategy that leverages a domain ontology to provide effective feedback. The concept hierarchy and co-occurrence between concepts in the domain ontology are drawn upon to ascertain partial correctness of a solution and guide student reasoning towards a correct solution. We describe the strategy incorporated in METEOR, a tutoring system for medical PBL, wherein the widely available UMLS is deployed and represented as the domain ontology. Evaluation of expert agreement with system generated hints on a 5-point likert scale resulted in an average score of 4.44 (Spearman's $\rho = 0.80$, $p < 0.01$). Hints containing partial correctness feedback scored significantly higher than those without it (Mann Whitney, $p < 0.001$). Hints produced by a human expert received an average score of 4.2 (Spearman's $\rho = 0.80$, $p < 0.01$).

1. Introduction

Problem-based learning (PBL) has become increasingly popular in medical schools as a means of training students and equipping them with the required clinical reasoning skills. A typical PBL session in the medical domain comprises a group of 6–8 students who work in collaboration to solve a given problem scenario [1]. Paying individual attention to a small group of students can place a heavy burden on faculty time, which is very costly. This is particularly true for medical faculty, who often have limited time to dedicate to teaching. Intelligent tutoring systems offer an attractive alternative in helping to train the students with the required clinical reasoning skills at no incremental cost per student.

Intelligent tutoring systems are interactive software applications that present a problem to the students in a particular domain. The students form their solution to the problem using the tutoring system interface. The system assesses the student solution and returns appropriate hints as feedback to guide the student towards a correct solution.

Tutoring systems normally contain either a set of approved solutions or, a mechanism that generates approved solutions to the problems presented to the students. Assessment of the student solution and feedback returned is tailored to be effective only within the knowledge confines of the approved solutions. Tutoring systems are typically unable to assess the partial correctness of student solutions when they fall outside the scope of the approved ones. Furthermore, for the purpose of solution representation, students are restricted to the choice of domain concepts from a custom built repository which is often quite narrow. Such characteristics lend themselves to a tutoring approach that is fairly brittle and quite opposed to how a human tutor would behave. A human tutor allows a diverse choice of domain concepts, assesses where the student solution lies in the broad knowledge space, acknowledges the partially correct aspects of the solution and guides the students back to the correct solution. Thus in order for a tutoring system to exhibit robust tutoring, it needs a broad knowledge base to allow students to explore a large space of solutions and work

* Corresponding author. Address: Isra University, 313 Hala Road, Hyderabad, 71000 Sindh, Pakistan. Fax: +92 22 2030185.
E-mail addresses: hkazi@isra.edu.pk (H. Kazi), haddawy@iist.unu.edu (P. Haddawy), ssiriwan@tu.ac.th (S. Suebnukarn).

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.02.010

creatively, while still being able to steer them towards a correct solution if they get off track.

An ontology presents great potential for reuse and as a knowledge base that could be exploited for reasoning purposes. Several tutoring systems have employed ontologies [2–4], but they require extensive effort in encoding the relevant knowledge into the ontology. The Constraint Acquisition System [5] uses a more feasible method of encoding the ontology constraints by learning from examples, but its initial design still needs to be defined manually.

The construction of a tutoring system typically requires knowledge acquisition in the three areas of domain model, student model and pedagogical model. Acquiring and encoding the relevant knowledge can lead to a large overhead in the development time of a tutoring system [6,7]. Attempts to expand the system and reuse the existing domain model for the rapid addition of new problems or cases are often hindered by the daunting task of acquiring the student model.

While the importance of the student model has been advocated [8], the design of some tutoring systems has excluded the student model based on the needs of the tutoring task [9]. Similar to Andes [9], our system too, does not use assessment to select the next task to be offered to the student. Because of the extensive effort required, tutoring systems often excel in one or two of the three models mentioned above and maintain a more simplified form of the remaining ones [10].

The development time for a tutoring system has also come under scrutiny in the comparison between Model Tracing (MT) and Constraint Based Modeling (CBM) [11,12]. Kodaganallur et al. [11] and Mitrovic et al. [12] have acknowledged the tradeoff between the reduction in development time and the quality of hints generated. The development time required to add a case is expected to vary based on the nature of the task domain. For the domain of statistical hypothesis testing, Kodaganallur et al. [11] report the development time of 5 person-days for problem modeling and 18 person-days for encoding the relevant knowledge in the case of CBM, whereas the development time was greater for MT. CBM simplifies the creation of new cases and has a reduced development time; however, its hints are not as effective and specialized as those in MT [11,12].

In order to ease the knowledge acquisition bottleneck, Martin and Mitrovic [13] adopt a CBM approach, where the student model is an overlay of the domain model constraints. Their student model simply contains a score of the times a constraint has been satisfied or violated during problem solving. However, defining and encoding the constraints is still a burdensome task. Defining the constraints would be even a greater burden and challenge for an ill-defined domain such as medical PBL [14].

In the domain of medical PBL, students may arrive at a solution from a variety of reasoning paths [15], making it a daunting task to build the student model. Based on our previous experience with the COMET system for medical PBL [16], it takes about one person-month to build the student model for each problem scenario. Modeling the diverse set of reasoning paths would be even more challenging if the system is expected to be robust in its tutoring approach by allowing students to explore a much broader solution space as mentioned above.

We extend our work on expanding the plausible solution space [15] by deploying the widely available knowledge source, the Unified Medical Language System (UMLS) [17], as the domain ontology in the METEOR tutoring system for medical PBL. In previous work [23] we had also presented a tool for authoring medical PBL cases using UMLS. In this paper we present a strategy for alleviating the overhead required to expand the tutoring system in adding new cases by omitting the student model. We exploit the structure of the domain ontology to assess the partial correctness of student solutions and generate hints that are relevant to the student activity during problem solving [30]. Furthermore, the time and effort required to add a new problem scenario to the tutoring system is also reduced.

2. Related work

2.1. UMLS in intelligent systems

The UMLS has been used for various purposes in the biomedical informatics domain, such as terminology development, lexical matching and biomedical document understanding. Qing and Cimino [31] extract knowledge of disease–chemical relationship from the UMLS for purposes of enriching electronic patient records for online perusal.

Mendonca and Cimino [26] describe work on extracting knowledge from MEDLINE citations for purposes of building a knowledge base. They analyze the search results to determine which semantic types are relevant to what kind of questions in Evidence Based Medicine, such as diagnosis, etiology, therapy and prognosis.

Achour et al. [28] describe a knowledge acquisition tool and how it could be employed 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 to use the semantic types and concepts in UMLS for reasoning purposes, but to use UMLS knowledge sources as a repository of terms from which a domain ontology could easily be constructed.

2.2. Semantic similarity

In order to provide students with partial correctness feedback, METEOR assesses the closeness of the student solution to a correct solution explicitly encoded into the system. This closeness is measured through the semantic similarity or semantic distance between relevant concepts.

Beginning with simple path length based measures [32,33] to advanced information theoretic metrics [34,35] researchers have developed methods through which, similarity between two concepts in an ontology, could be defined in quantitative terms. Most similarity measures determine the lowest common subsumer (LCS) of the two concepts, to compute the path length from one node to the other node through this LCS. The LCS is the lowest node in the hierarchy that is a common ancestor to both the nodes, between which semantic distance is to be measured.

There has been growing interest in defining and applying measures of semantic distance, for medical terminologies and the UMLS. Caviedes et al. [27] develop a quantitative metric that can enable intelligent systems to differentiate between concepts in UMLS and measure their semantic distance. They describe their results for PAR (parent–child) links between concepts based on three terminologies within UMLS, MeSH, SNOMED-CT and ICD9CM. They adopt a simple edge counting procedure to compute the conceptual distance between two concepts over the shortest path between them, while simply mentioning the depth of the concepts in the hierarchy, as a possible influencing factor in the similarity measure.

Al-Mubaid and Nguyen [22] present an information theoretic approach to compute the semantic distance between two given concepts in an ontology. They use a cluster-based approach where the depth of the tree cluster, containing the relevant concept nodes is used along with a scaled measure of the path length between respective concept nodes. Concepts that lie deeper in the ontology tree will be more similar based on the specificity of information.

Pedersen et al. [25] discuss and analyze a set of existing semantic similarity measures and describe a context vector measure based on medical corpora. They compare the context vector measure with existing measures as applied to a commonly used dataset of

concept pairs. They found that the context vector measure was as effective as the ontology-based measures provided the corpus was large enough.

Batet et al. [24] present and compare their approach with other semantic similarity measures as applied to SNOMED-CT, using a dataset commonly used for evaluation of semantic similarity measures in the biomedical domain. They employ a technique which does not take into account the specificity of the information content. Their measure is based on the ratio between the amount of non-shared knowledge and the sum of shared and non-shared knowledge between the relevant concepts.

2.3. Intelligent tutoring systems

Use of UMLS has also found its way through intelligent tutoring systems into the medical domain. Crowley et al. [20] 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. The system does not make use of an explicit student model. Their work is similar to ours in generating hints without differentiating between two students having different knowledge levels performing the same exercise.

The Docs 'n Drugs tutoring system [36] 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.

While the UMLS has been used for different purposes in various applications, to the best of our knowledge, UMLS has not been previously used as the main knowledge source for inference or reasoning purposes in an intelligent tutoring system. The concept of partial correctness has been discussed in the context of tutoring systems [4,19], wherein a part of the solution is explicitly recognized as correct. Our notion of partial correctness is different and is assessed through knowledge inference rather than explicitly encoded knowledge. Fiedler and Tsovaltzi [19] employ a domain ontology for tutoring mathematics theorem proving. The domain ontology of concepts contains some objects and relations defined as anchoring points, which serve as the basis for the content of the generated hints. Our hint generation strategy is different and draws inferences from the structure of the existing domain ontology at run-time without recourse to explicit encoding of knowledge into the ontology.

The design 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, to form their solution. The CIRCSIM-Tutor [4] teaches cardiovascular physiology by describing a perturbation of a cardiovascular condition, and initiating a question and answer dialog with the student. 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 [2] teaches dermatopathology by presenting a visual slide as a problem scenario and asks students to classify the diseases. Solutions accepted by the tutoring system are based on the ontology customized for the system. Thus students are not allowed to present alternative plausible hypotheses that may lie beyond the scope of the customized ontology.

3. Medical PBL and system prototype

In a typical PBL session in the medical domain, a problem scenario is presented to a group of 6–8 students, who form their hypothesis in the form of a causal graph, 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 [21]. Enabling conditions are factors that trigger the onset of a medical condition, e.g., aging, smoking; faults are the bodily malfunctions that result in various signs and symptoms, e.g., pneumonia, diabetes; consequences are the signs and symptoms that occur as a result of the diseases or disorders, e.g., fatigue, coughing.

Our work derives from the COMET system [16] designed to cover medical PBL for various domains. In COMET each problem scenario is first referred to human domain experts who provide an expert solution that is eventually encoded into the system. Student solutions are compared against this expert solution for evaluation. Thus a plausible student solution that does not match the expert solution is not entertained. The system allows students to form their hypothesis by choosing medical concepts from a repository manually encoded into the system. Students are given feedback based on the current state of their knowledge, which is assessed against a student model [16].

In our new system Medical Tutor Employing Ontology for Robustness (METEOR), problem solutions collected from experts are combined with UMLS tables to form the domain model. The pedagogical module of the system comprises a hint generation mechanism that leverages off of the UMLS concept hierarchy and provides students a measure of partial correctness of their hypotheses. Assessment of student solutions is not used to select the next step or task to be offered to the students. Furthermore, the hint generation employs the rich domain knowledge of the UMLS in lieu of a student model. Thus the design of our tutoring system does not include a student model.

The problem representation in METEOR is the same as that in COMET of a directed acyclic graph for forming the hypothesis. The student user is provided with a workspace as a hypothesis board to form the hypothesis, along with a text chat pane that returns hints to guide the student in clinical reasoning, as shown in Fig. 1. The student chooses concepts from the UMLS Metathesaurus [17] as hypothesis nodes and draws edges between nodes, using a mouse. The problem solving activity begins as the student is presented a problem scenario, such as the one shown in Fig. 1.

After studying the above problem scenario related to diabetes, the student hypothesizes that *Diabetes Mellitus* is a cause of *Hyperglycemia*, which is shown to be a cause of *Diabetic Neuropathy*, as shown in Fig. 1.

4. System domain model

The UMLS [17] is a widely available medical knowledge source and is essentially a collation of various medical ontologies and terminologies (MeSH, SNOMED-CT, Gene Ontology, etc.). The broad and diverse UMLS contains about two million medical concepts covering various medical domains [17].

The system domain model comprises 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 causal relationships between various medical concepts, such as:

Hyperglycemia → decreased glucose transport into cells
Diabetic neuropathies → numbness
Decreased glucose transport into cells → fatigue

The expert knowledge base is formed through the collation of expert solutions to various problem scenarios. On average each

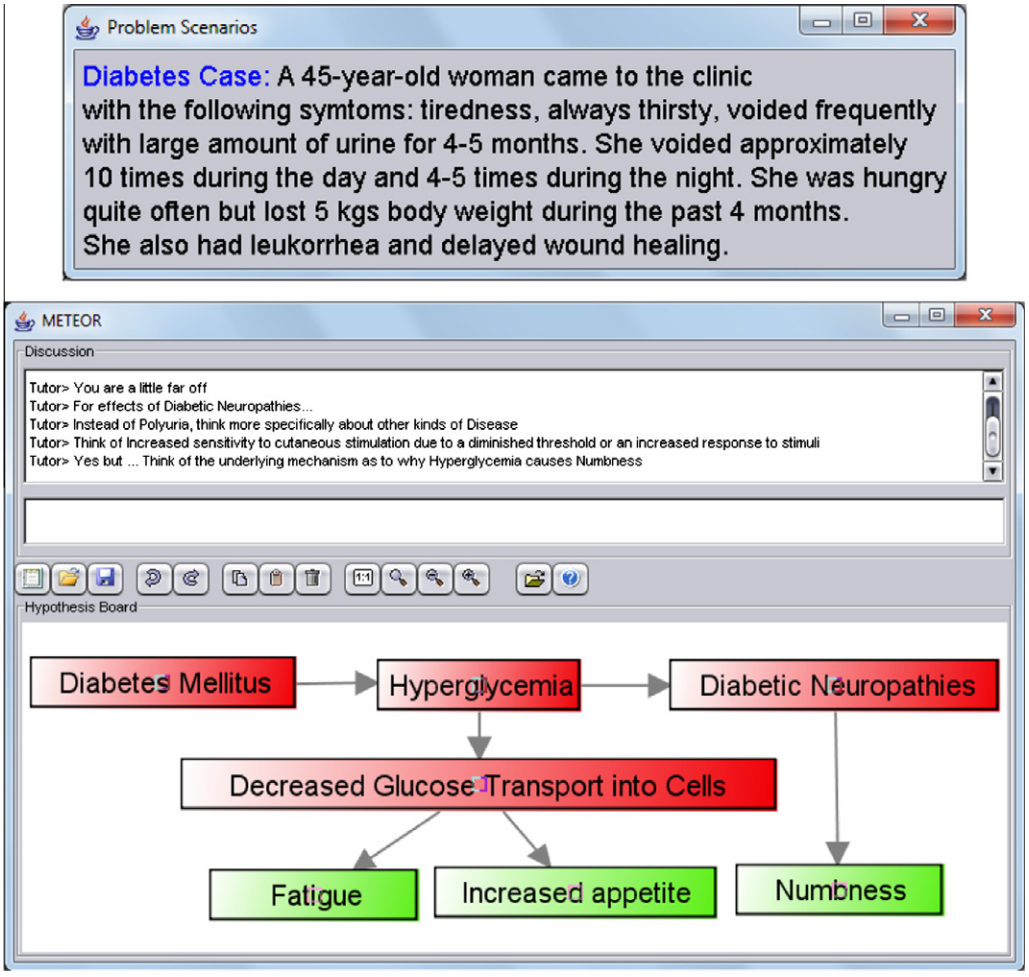

Fig. 1. System prototype interface.

expert solution leads to the addition of about 70–80 causal links to the expert knowledge base. The construction of an expert solution requires about 3–4 h. 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.

5. Pedagogy of assessment and feedback

The hints generated by the system are composed of two elements: assessment of the partial correctness of the student solution and guidance towards a correct solution. Each hypothesis causal link drawn by the student is evaluated by the system through a strategy that accepts plausible solutions beyond the scope of the explicitly encoded ones [18]. 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 as feedback to the student. If the causal relationship drawn by the student is essentially correct but requires additional intermediate nodes in between, then the system disallows the edge to be drawn and encourages the student to describe the underlying mechanism. For example, considering the diabetes case described above, if the student draws the link: *hyperglycemia → numbness*, the system responds with the hint: **"Yes, but ... Think of the underlying mechanism as to why *hyperglycemia* causes *numbness*."** On the other hand, if the student draws the reverse link: *numbness → hyperglycemia*, the system responds with the hint: **"On the contrary, think of hyperglycemia as a cause of numbness."**

If the student link does not fall into any of the cases described above, the system makes use of a heuristic method to assess its partial correctness and deliver a hint to guide the student towards a correct link. The purpose of partial correctness feedback is to inform the student how close his/her solution is to be accepted. The hint pre-amble containing the partial correctness feedback is phrased as one of the following: (1) "You are very close", (2) "You are somewhat close", (3) "You are a little far off", (4) "You are quite far off", (5) "Hmm... Not sure. They may be a causal relation between the two", and (6) "Hmm... Can't say about the relation between the two."

5.1. Example 1: partial correctness through semantic distance

Imagine a situation related to the diabetes case mentioned above, where a student tries to draw a causal link: *hyperlipidemia → diabetic neuropathy*. Suppose the *expert knowledge base* does not recognize this link, however it recognizes that there is an expert link: *hyperglycemia → diabetic neuropathy*. In other words, what should have been *hyperglycemia* has been hypothesized by the student to be *hyperlipidemia*.

In order to assess the partial correctness of the student link, the system tries to find the semantic distance between *hyperlipidemia* and *hyperglycemia*. The semantic distance is measured by employ-

ing a modified version of the method described by Al-Mubaid and Nguyen [22].

Al-Mubaid and Nguyen [22] compare their method with other semantic similarity measures in the context of the biomedical domain by applying them against different terminologies in the UMLS and report better results. Furthermore, their semantic similarity measure allows the flexibility to limit the search space in the concept hierarchy and return results in lesser time compared to a full scan of the hierarchy, in exchange for reduced accuracy. For the purpose of an interactive tutoring system, the system response time is crucial in maintaining student motivation in performing the given exercise. Based on our experimental results, both in terms of speed and accuracy, we decided to employ the semantic similarity measure of Al-Mubaid and Nguyen [22] and modify it by limiting the search space.

The semantic distance is computed in the following manner:

$$SemDist(a,b) = ln((PathLength - 1)^{\alpha} * ComSpec^{\beta} + k)$$

where *PathLength* is the number of nodes traversed from a to b, α, β and k are tuning parameters, and *ComSpec* is the common specificity of nodes a and b, computed as shown below:

$$ComSpec(a,b) = DepthOfCluster - \text{depth}(LCS(a,b))$$

where *DepthOfCluster* is the depth of the cluster which contains the LCS node, α and β are tuning variables and k is a constant. The LCS node is the lowest common subsumer or the lowest ancestor that is common to both the concept nodes, between which semantic distance is computed. As shown in Fig. 2a, the node *metabolic diseases* is the LCS of nodes a and b, where node a is *hyperlipidemia* and b is *hyperglycemia*. The depth of the LCS node is one, whereas the depth of the cluster containing the LCS node, is three. The path length between nodes a and b is four, by node counting.

For example, for the structure in Fig. 2a, and for the tuning parameter values of $\alpha = 3$, $\beta = 1$ and $k = 1$, the semantic distance between *hyperlipidemia* and *hyperglycemia* is 4.0073.

Parent–child relationships from the UMLS Metathesaurus are used to construct the hierarchy of both nodes between which semantic distance is to be measured, as shown in Fig. 2a. Based on the value of the semantic distance, the system classifies whether the nodes are *very close, somewhat close, a little far,* or *quite far*. Based on our experiments and feedback from domain experts, we defined the following thresholds for semantic distances: less than 1.8 (very close), between 1.8 and 4.5 (somewhat close), between 4.5 and 8 (a little far off), and greater than 8 (quite far). The thresholds are an outcome of experiments performed in our previous work in ascertaining the acceptability of a causal link drawn by a student [15]. The thresholds were calibrated against the human tutor ratings awarded to the acceptability of causal links drawn by students. For semantic distances less than 1.8, causal links were found to be rated as acceptable or close to acceptable, while for those greater than 8, the acceptability rating was too low. The remaining two intervals were evenly spaced between these cutoffs. The hint wordings such as *very close* and *somewhat close* were framed based on input from human tutors.

5.1.1. Guidance towards the correct solution

In order to guide the student towards a correct solution, the system examines the parent–child hierarchy to judge the commonality between the student link and a correct expert link. The system tries to find the lowest node in the hierarchy that is a common ancestor to both concepts in question: *hyperlipidemia* and *hyperglycemia*. The system finds that *metabolic diseases* is a common ancestor to both the concepts, as shown in Fig. 2a. Thus the system infers that the student knows that a kind of *metabolic disease* leads to *diabetic neuropathy*, however the student is not clear which kind. The hint content is framed to guide the student reasoning from its current position to the correct solution. This reasoning path of the hint content is shown in the dotted arrow in Fig. 2a, which leads from *hyperlipidemia* round the common ancestor towards *hyperglycemia*. Based on the assessment of partial correctness and the reasoning path en route the correct solution, the system responds with the hint: **"You are somewhat close. For causes of *diabetic neuropathy* ... Instead of *hyperlipidemia*, think about other kinds of *metabolic diseases*. Think of *A heterogeneous group of disorders characterized by glucose intolerance*".**

Here, 'A heterogeneous group of disorders characterized by glucose intolerance' is the definition in UMLS for the concept: *glucose metabolism disorder*. In other words the system gives the hint template: "Instead of ⟨student node⟩, think about other kinds of ⟨common ancestor⟩ and ⟨definition of next child in line from the common ancestor towards the expert node⟩".

If the student draws the link, *renal glomerular disease → diabetic neuropathy*, the system measures the semantic distance between *hyperglycemia* and *renal glomerular disease* and finds the distance as 5.52, thus the two nodes to be *a little far off*. The hint is framed: **"You are a little far off. For causes of *diabetic neuropathy* ... Instead of *renal glomerular disease*, think more specifically about other kinds of *Disorder of body system*. Think of *Abnormally high BLOOD GLUCOSE level, beyond the normal range*."**

If the student draws the link *glucose metabolism disorder → diabetic neuropathy*, the system measures the semantic distance between *hyperglycemia* and *glucose metabolism disorder* and finds the two nodes to be *very close* and accepts the student link by giving the hint: **"You are very close. I was thinking of *hyperglycemia → diabetic neuropathy*, but *glucose metabolism disorder* is also acceptable. Good."**

5.2. Example 2: partial correctness through co-occurrence frequency

Imagine a situation related to a heart attack case, where a student tries to draw a causal link: *hyperlipidemia → hyperglycemia*. The system does not find this link to be acceptable, however, it finds an expert causal link: *hyperlipidemia → endothelial degeneration*. In other words what should have been *endothelial degeneration* has been hypothesized by the student to be *hyperglycemia*.

The system tries to find a common ancestor to both *hyperglycemia* and *endothelial degeneration*, but is unable to find one. In this situation, the system cannot assess the partial correctness through the semantic distance measure. As a weaker measure, it checks to

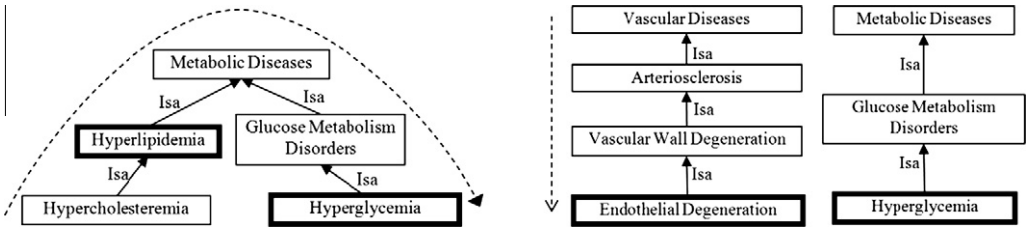

Fig. 2. (a) Concept hierarchy: example 1. (b) Concept hierarchy: example 2.

Expert Link	Student Link
Lung Consolidation → Pneumonia	Pulmonary Emphysema → Pneumonia

Hint From Tutor A

However, you are somewhat close

For causes of Pneumonia...

Instead of Pulmonary Emphysema, think more specifically about other kinds of Lesion of lung

5 · 4 · 3 · 2 · 1

Hint from Tutor B

For causes of Pneumonia...

Instead of Pulmonary Emphysema, think more specifically about other kinds of Lesion of lung

5 · 4 · 3 · 2 · 1

Correct Link	Student Link
Coronary Arteriosclerosis → Chest pain	Right heart failure → Chest pain

Hint From Tutor A

Tutor> However, you are somewhat close
Tutor> For causes of Chest Pain...
Tutor> Instead of Right heart failure, think more specifically about other kinds of Heart Diseases
Tutor> Think of Thickening and loss of elasticity of the coronary arteries, leading to progressive insufficiency of the arteries (CORONARY DISEASE)
5 · 4 · 3 · 2 · 1

Hint from Tutor B

Tutor> For causes of Chest Pain...
Tutor> Instead of Right heart failure, think more specifically about other kinds of Heart Diseases
Tutor> Think of Thickening and loss of elasticity of the coronary arteries, leading to progressive insufficiency of the arteries (CORONARY DISEASE)
5 · 4 · 3 · 2 · 1

Fig. 3. Samples of hint for evaluation.

see if the UMLS has any information regarding the co-occurrence of *hyperlipidemia* and *hyperglycemia* in medline citations. If the normalized co-occurrence frequency is found to be greater than zero, the system forms the hint pre-amble: **"Hmm... There may be a causal relation between hyperlipidemia and hyperglycemia."** Otherwise the following hint pre-amble is formed: **"Hmm... Can't say about the causal relation between *hyperglycemia* and *hyperlipidemia*."**

5.2.1. Guidance towards the correct solution

In order to guide the student towards the correct solution, the system adopts an approach similar to the one described for Example 1. Since there is no common ancestor in this case, the system tries to direct the student towards *endothelial degeneration* by starting from a few ancestors above, as shown in Fig. 2b. The hint is framed as: **"For effect of *hyperlipidemia* ... Instead of *hyperglycemia*, think of kinds of *vascular diseases* and *thickening and loss of elasticity of arterial walls*."**

Here, 'Thickening and loss of elasticity of arterial walls' is the definition in UMLS for the concept *arteriosclerosis*. In other words the system gives the hint template: "Instead of ⟨student node⟩, think about kinds of ⟨great grandfather of expert node⟩ and ⟨definition of grandfather of expert node⟩".

In both COMET and METEOR, hints are generated to guide the student towards a particular causal path, such as: (1) *Intracranial Pressure Increase → Brain Damage → Unconsciousness* and (2) *Lung Consolidation → Pneumonia*. In the COMET system [29], the causal path for hint generation is selected based on the probabilistic student model. The student model is in the form of a Bayesian network, which contains the probabilities of the various candidate causal paths that indicate the likelihood of the student thinking along the lines of the relevant path. Thus, COMET offers hints based on the current cognitive state of the student. In METEOR [30], the causal path for hint generation is selected based on a heuristic measure of the closest candidate causal path, which is estimated using the semantic distance measure described above. Thus the

hint generation strategy is probabilistic in COMET and heuristic in METEOR.

6. Evaluation

We classified the hints into different kinds as described in section 5, such as: 1. "You are very close", 2. "You are somewhat close". We then collected system generated hints from student log files. In order to enable stratified sampling, these hints were then randomly selected from each class of hints, so each class would have an even representation in a total sample of 30. In order to gauge human agreement with the system generated hints, we conducted separate evaluations with both experts and students. In the expert evaluation, five faculty members from Thammasat University having more than five years of experience in using PBL in teaching medicine, were asked to rate the sample of hints on a 5-point likert scale: 1 (strongly disagree) to 5 (strongly agree). For each sample, experts were shown the causal link drawn by the student and the corresponding expert link as a correct solution, along with the hint generated by the system, as shown in Fig. 3. In order to evaluate the utility of the partial correctness feedback, experts were presented two versions of the same hint. As shown in Fig. 3a, hint from Tutor A contains the pre-amble of partial correctness feedback, for example 'You are somewhat close', whereas the Tutor B hint is without this feedback pre-amble.

Hints containing partial correctness led to an average score of 4.44 (Spearman's $\rho = 0.80$, $p < 0.01$), whereas those without it led to an average score of 3.58 (Spearman's $\rho = 0.78$, $p < 0.01$). Hints with partial correctness scored significantly higher than those without it (Mann Whitney, $p < 0.001$). The average of ratings awarded by the five experts to hints generated by the system are shown in Fig. 4.

In the second evaluation we had the same sample of hints rated by ten medical students of second year. Hints containing partial correctness led to an average score of 4.62 (Spearman's $\rho = 0.82$, $p < 0.01$), whereas those without it led to an average score of 3.78 (Spearman's $\rho = 0.71$, $p < 0.01$). One student awarded a rating score of 5 to all hints with partial correctness; these ratings were hence not used in computing the statistical agreement. Hints with partial correctness scored significantly higher than those without it (Mann Whitney, $p < 0.001$). The average of ratings awarded by the ten students to system generated hints are shown in Fig. 5.

In order to compare the quality of system generated hints against a gold standard, we conducted a separate evaluation by asking a human expert to provide hints on the same set of scenarios. These hints were then presented to five faculty members from Thammasat University for rating, to establish a gold standard. The average rating score turned out to be 4.20 (Spearman's $\rho = 0.80$, $p < 0.01$). A statistically significant difference was not found between these scores and those obtained from the expert evaluation of system generated hints with partial correctness (Mann Whitney, $p = 0.287$). The average of ratings awarded by five experts to hints received from the expert are shown in Fig. 6.

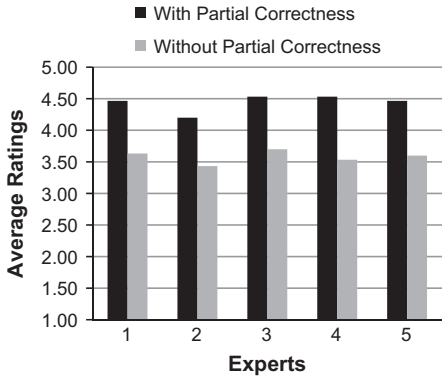

Fig. 4. Rating of system generated hints by five experts.

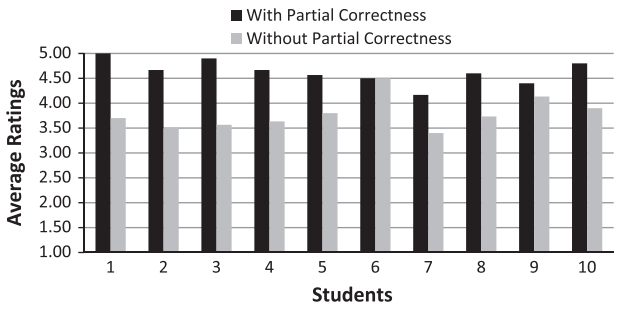

Fig. 5. Rating of system generated hints by ten students.

7. Discussion

The expert rating of system generated hints with partial correctness was found to be comparable to the gold standard, since a statistically significant difference was not found between the two groups. The overall average expert and student rating of 4.44 and 4.62, indicate strong expert and student acceptance of the system generated hints. Hints including the element of partial correctness scored significantly higher than those without it, which shows that both experts and students found the partial correctness feedback to be very useful.

The ratings by students were almost the same as those awarded by experts, with the only exception of two hints. These two hints were rated low by experts because they found them to be too direct and revealing the answer. Understandably the students thought different and were not as critical of those hints, though they still rated these two hints lower than the others.

According to one PBL expert, some of the content in the sample of hints was even better than what an average PBL tutor would be able to formulate. This is because not all PBL tutors are expert in all of the PBL cases. Their knowledge about concepts is sometimes lacking in certain areas and they are not always able to formulate the right description for a particular concept. In fact, such problems with hint quality due to knowledge gaps occur in METEOR as well when the definition text is missing for concepts in UMLS. It is worth noting that hints that contained the concept definition text scored higher than those where this text was missing in UMLS.

Inference techniques applied to a large knowledge source such as UMLS, can be quite taxing on the processing power and result in

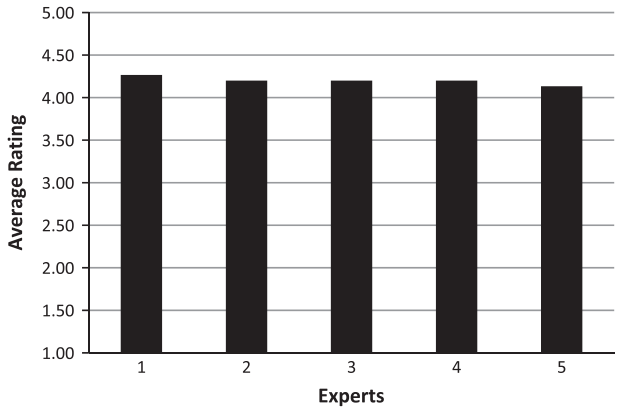

Fig. 6. Rating of expert generated hints by five experts.

delayed system response. To address this problem, we reduced the search space, but this can lead to the reduced quality of hints since relevant concepts in the student and expert solution may not be connected in the hierarchy formed through the reduced search space. In this case, the partial correctness through semantic distance would not be assessed and the system would resort to the co-occurrence frequency. The hint guiding the student towards a correct solution will also be framed accordingly, devoid of the common ancestor to both concepts. However, considering that the relevant nodes are quite far apart, this may not have significant impact on the quality of the hint.

The semantic distance between parent and child concepts is not consistent throughout UMLS. In some cases the parent concept may be semantically close to its child concept, whereas in other cases they may be semantically quite distant. The parent concept may be too broad compared to its child concept or too vague; thus leading to a hint that is less relevant.

8. Conclusions

In this paper we have described how to ease the bottleneck of expanding a tutoring system. We have described how an existing broad knowledge source such as the UMLS, can be deployed as the domain ontology and its structure leveraged to assess the partial correctness of the student solution and generate hints based on the context of the student activity. Compared to the previous version of COMET, the time for the development and encoding of a new problem scenario has been drastically reduced from one person-month to 4–5 person-hours.

We have described the system implementation in the context of medical PBL, but the techniques could easily be applied to other domains where the task involves causal relationships and the domain ontology also contains a textual definition of the concepts. The techniques could be particularly relevant for other ill-defined domains, which require greater flexibility in assessment and feedback.

In interpreting the results of the proposed techniques, it is worth noting that the domain ontology has not been crafted specially for the task of medical PBL. A purpose built domain ontology is likely to yield better results, especially when its utility for hint generation is considered at the time of design.

The UMLS is incrementally expandable and a generic knowledge source for the broad biomedical domain, which is developed and distributed at a very wide scale. While existing versions of UMLS have limited support, later versions may feature support for a greater variety of tasks such as accurately inferring causal relations between concepts. Such additions will greatly facilitate the effective deployment of UMLS as a knowledge source for intelligent applications. Tutoring systems such as METEOR, would then be able to allow students to be more creative and explore novel solutions to problems. METEOR would then, also be more accurate in providing assessment and feedback.

9. Limitations and future work

One limitation of the study is the small sample size of 30 causal links against which students and experts were asked to rate their acceptance. Also, there may be inter subject clustering, which will have an impact on the statistical power, especially in the evaluations performed by experts limited to five. In the evaluations performed by 10 students, reduced effective sample sizes still lead to significant statistical agreements.

Furthermore, our hint generation strategy leveraging an existing knowledge source does not take into account the possibility of students having misconceptions at the ontology level. This could be addressed in a future study.

For future work, we would like to evaluate the hint generation for cases other than the well trod areas of heart attack, diabetes and pneumonia. Although the UMLS is large enough to cover virtually any problem scenario in the broad medical domain, this may impact the quality of hints, if there are knowledge gaps in UMLS in those areas. We would also like to investigate the impact of different hints on the student learning outcomes.

Finally we would like to compare and examine the tradeoffs between the nature of clinical reasoning gains acquired through METEOR and through COMET, especially in light of the fact that as previous studies have shown, a feedback strategy such as the one proposed in this paper, may not be as effective as those that stem from a carefully captured cognitive student model. Nonetheless the tradeoff may be worth it, if one considers the long term ramifications in adding new cases for the large scale deployment of tutoring systems for instructional purposes.

Acknowledgments

We would like to thank the students and general practitioners at Thammasat University for their time and effort during the evaluations.

References

[1] Barrows H. A taxonomy of problem-based learning methods. Med Educ 1986;20:481–6.
[2] Crowley R, Medvedeva O. An intelligent tutoring system for visual classification problem solving. Artif Intell Med 2006;36(1):85–117.
[3] Day MY, Lu C, Yang JD, Chiou G, Ong CS, Hsu W. Designing an ontology-based intelligent tutoring agent with instant messaging. In: Fifth IEEE international conference on advanced learning technologies; 2005. p. 318–20.
[4] Mills B, Evens M, Freedman R. Implementing directed lines of reasoning in an intelligent tutoring system using the atlas planning environment. In: International conference on information technology; 2004. p. 729–33.
[5] Suraweera P, Mitrovic A, Martin B. A knowledge acquisition system for constraint based intelligent tutoring systems. In: Conference on AI in, education; 2005. p. 638–45.
[6] Anderson JR, Corbett A, Koedinger K, Pelletier R. Cognitive tutors: lessons learned. J Learn Sci 1996;4(2):167–207.
[7] Mitrovic A. Experiences in implementing constraint-based modelling in SQL-Tutor. In: 4th International conference on intelligent tutoring systems; 1998. p. 414–23.
[8] Self JA. Bypassing, the intractable problem of student modelling. In: Proceedings of ITS 1988, Montreal; 1988. p. 18–24.
[9] VanLehn K, Lynch C, Schulze K, Shapiro JA, Shelby R, Taylor L, et al. The Andes physics tutoring system: lessons learned. Int J Artif Intell Educ 2005;15:147–204.
[10] Murray T. Expanding the knowledge acquisition bottleneck for intelligent tutoring systems. Int J Artif Intell Educ 1997;8:222–32.
[11] Kodaganallur Viswanathan, Weitz Rob , Rosenthal David. A comparison of model-tracing and constraint-based intelligent tutoring paradigms. Int J Artif Intell Educ 2005;15(2):117–44.
[12] Mitrovic A, Koedinger KR, Martin B. A comparative analysis of cognitive tutoring and constraint-based modeling. Johnstown, PA, USA: User Modeling 2003: 9th International Conference (UM 2003); 22–26 June 2003. Lecture Notes in Computer Science, 2702. p. 313–22.
[13] Martin B, Mitrovic A. Easing the ITS bottleneck with constraint-based modelling. New Zealand J Comput 2001;8(3):38–47.
[14] Pople Jr HE. Heuristic methods for imposing structure on Ill-structured problems: the structuring of medical diagnostics. In: Szolovits P, editor. Artificial intelligence in medicine. Boulder (Colorado): Westview Press; 1982 [chapter 5].
[15] Kazi H, Haddawy P, Suebnukarn S. Expanding the space of plausible solutions in a medical tutoring system for problem based learning. Int J Artif Intell Educ 2009;19(3):309–34.
[16] Suebnukarn S, Haddawy P. Modeling individual and collaborative problem-solving in medical problem-based learning. User Model User-Adap Inter 2006;16(3–4):211–48.
[17] US National Library of Medicine, <http://www.nlm.nih.gov/research/umls/>.
[18] Kazi H, Haddawy P, Suebnukarn S. Expanding the space of plausible solutions for robustness in an intelligent tutoring system. In: Proceedings of 9th international conference on intelligent tutoring systems, Montreal, Canada; 2008. p. 583–92.
[19] Fiedler A, Tsovaltzi, D. Domain-knowledge manipulation for dialogue-adaptive hinting. In: Proceedings of the 12th international conference on artificial intelligence in education (AIED 2005); 2005. p. 801–3.
[20] Crowley RS, Tseytlin E, Jukic D. ReportTutor – an intelligent tutoring system that uses a natural language interface. In: Proc AMIA symp 2005. p. 171–5.

[21] Feltovich PJ, Barrows HS. Issues of generality in medical problem solving. In: Schmidt HG, De Volder ML, editors. Tutorials in problem-based learning: a new direction in teaching the health professions. The Netherlands: Van Gorcum; 1984.
[22] Al-Mubaid H, Nguyen HA. A cluster based approach for semantic similarity in the biomedical domain. In: Proceedings of the 28th IEEE EMBS annual international conference, New York, USA; August 30–September 3, 2006.
[23] Suebnukarn S, Haddawy P, Rhienmora P. A collaborative medical case authoring environment based on the UMLS. J Biomed Inform 2008;14(2):318–26.
[24] Batet M et al. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 2011;44(1):118–25.
[25] Pedersen Ted, Pakhomov Serguei VS, Patwardhan Siddharth, Chute Christopher G. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40(3):288–99. ISSN 1532-0464.
[26] Mendonca EA, Cimino JJ. Automated knowledge extraction from MEDLINE citations. In: Proceedings of AMIA 2000 fall, symposium; 2000. p. 575–9.
[27] Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric. J Biomed Inform 2004;37(2004):77–85.
[28] Achour SL, Dojat M, Rieux C, Bierling P, Lepage E. A UMLS-based knowledge acquisition tool for rule-based clinical decision support systems development. J Am Med Inform Assoc 2001;8(4):351–60.
[29] Suebnukarn S, Haddawy P. A Bayesian approach to generating tutorial hints in a collaborative medical problem-based learning system. Artif Intell Med 2006;38(1):5–24.
[30] Kazi H, Haddawy P, Suebnukarn S. Leveraging a domain ontology to increase the quality of feedback in an intelligent tutoring system. In: Proceedings of the 10th international conference on intelligent tutoring systems, Pittsburgh, USA; 2010. p. 75–84.
[31] Qing ZMS, Cimino JJ. Automated Knowledge Extraction from the UMLS. In: Proceedings of the 1998 AMIA Annual Fall, Symposium; 1998. p. 568–72.
[32] Rada R, Mili H, Bicknell E, Blettner M. Development and application of a metric on semantic nets. IEEE Trans Syst Man Cybernet 1989;19(1):17–30.
[33] Leacock C, Chodorow M. Combining local context and wordnet similarity for word sense identification. In: Fellbaum C, editor. WordNet: an electronic lexical database, 1998. Cambridge: MIT Press; 1998. p. 265–83.
[34] Resnik P. Semantic similarity in a taxonomy: an information based measure and its application of problems of ambiguity in natural language. J Artif Intell Res 1999;11:95–130.
[35] Lin D. An information theoretic definition of similarity. In: Proceedings of the 15th international conference on, machine learning; 1998. p. 296–304.
[36] Martens A, Bernauer J, Illmann T, Seitz A. Docs 'n drugs – the virtual polyclinic. An intelligent tutoring system for web-based and case-oriented training in medicine. In: Proceedings of the AMIA fall symposium, 2001.