Accepting Inferred Student Solutions by Tutoring System in an Ill-Defined Domain

Hameedullah Kazi
Isra University, Hyderabad, Pakistan

Asia Kainat Awan
Isra University, Hyderabad, Pakistan

ABSTRACT

Intelligent Tutoring Systems have made great advances in providing assessment and useful feedback in domains with well-structured problems, where start state, rules, or goals of a problem are well formalized and used to reach an unambiguously correct or incorrect solution. The problems of ill-defined domain often possess multiple solutions. Plausible student solutions of ill-defined problems are deemed wrong by tutoring system if they do not match the known solution accepted by the system. This paper describes a mechanism and the results of a tutoring system in an ill-defined domain such as the English language, for accepting plausible student solutions for ill-defined problems. The WordNet is deployed as a knowledge base, which is a lexical resource of English language database. Semantic similarity measure technique uses WordNet ontology hierarchy to accept the student plausible solutions. The student solutions of cloze passages were evaluated by a group of English experts and compared against a semantic similarity measure. The experts agreed among themselves with a correlation of 0.7 with p<0.05. The correlation between semantic similarity and experts is 0.58 with p<0.05 to indicate valid hypothesis. The **area under the curve of ROC is 0.76.**

Index Terms

Tutoring system, ill-defined domain, WordNet, robustness, plausible solution

1. INTRODUCTION

Intelligent tutoring system (ITS) is any computer-based learning system which attempt to provide direct customized instruction or feedback to student, while performing a task. Intelligent Tutoring Systems have made great strides in providing assessment and useful feedback in well-defined domain.

Well-defined domains tutors are characterized by a basic formal theory, strong domain model to structure the relevant domain knowledge and to validate student actions. Operationalizing the domain theory makes it possible to identify study problems, provide a clear problem solving strategy, and assess student solutions definitively based on the existence of unambiguous answers which is classified as correct or incorrect [1]. Well-defined domain tutors can provide help readily by comparing student problem solving steps to the existing domain model. This has permitted the development of strong, domain-general methodologies such as model-tracing systems and constraint-based tutors [2]. Examples include well-defined domain tutors for Physics (Andes Tutor) [3], Mathematics (ActiveMath) [4], Algebra (PAT) [5], AGP [3], Database (SQL-Tutor) [6].

Ill-defined domains often depend on reasoned argument rather than formal proofs [2]. Ill-defined problem has multiple solutions and there exists no unambiguously correct solution which makes the process of assessing student solution, providing feedback, modeling student knowledge, or measuring performance correspondingly a little difficult.

Tutoring systems are normally constructed with some problem scenarios together with their solutions, which have been pre-approved by human experts. In some cases the solutions are built in to the system, while in other cases they are produced on run time based on logical rules. If a student solution coincides with one stored into or generated by the system, then it is deemed acceptable. However, if the student solution is otherwise correct but not the same as the one recognized by the system, then it is rejected as incorrect. As a result of this, students who use the tutoring system, train themselves to produce just the solutions accepted by the system and they are not encouraged to think creatively and explore a wider variety of solutions [7].

There have been several ITSs developed for English domain, such as English Tutor [8], the VP2 [9], Spengels [10], Compounds [11] for learning to use compounds in English, CAPIT [12] teaches capitalisation and punctuation, Passive Voice Tutor [13] teaches passive voice of English grammar and the REAP [14, 15] for learning vocabulary, reading comprehension. However, existing tutoring systems of English have not focused on providing a greater solution space than the one explicitly encoded into the system. This paper provides a mechanism of expanding the solution space and reports its accuracy results.

2. METHODS

Examples of identified problems leading to ill-definedness are English cloze passages. A cloze passage is a piece of text in which words have been omitted throughout. The objective of cloze activity is to increase reading comprehension. The objective for the student is to predict words that belong in the blanks of the cloze passage. There could be multiple acceptable students predict words for each omitted words. Passage context makes a little easier for students to recognize the correct solution in a cloze passage.

We have considered only nouns and verbs to be omitted throughout in cloze passages. The two cloze passages [19] have been changed by filling up those omitted words which were prepositions, adjectives and adverbs and then omitting some words which were nouns and verbs. One cloze passage was just a passage, but was made it a cloze passage by omitting some words which were nouns and verbs. One example of cloze passage is given below.

A. Example of Cloze Passage

"When all the people had assembled, the king surrounded by his court, ____1_____ a signal. Then a door beneath him opened, and the accused man stepped out into the arena. Directly opposite him were two doors, exactly alike and side by side. It was the duty and the privilege of the person on trial to walk directly to these doors and open one of them. He could open either door he ____2_____; he was subject to no guidance or influence but that of impartial and incorruptible chance. If he opened the one, there ____3____ of it a hungry tiger, the fiercest and most cruel that could be found, which immediately sprang upon him and ____4_____ him to pieces as a punishment for his guilt. But, if the accused person opened the other door, out of it came a beautiful lady, and to this lady he was immediately married, as a ____5_____ of his innocence. This was the king's ____6_____ of administering justice. Its perfect fairness is obvious. The criminal could not know out of which door would come the lady; he opened either he pleased, without having the slightest ____7_____ whether, in the next instant, he was to be devoured or ____8_____. So the accused person was instantly ____9_____ if guilty, and, if innocent, he was rewarded on the spot". Adapted from The Lady or the Tiger by Frank Stockton [19].

B. WordNet

To represent English language knowledge in an ITS domain model, WordNet database is employed as a domain knowledge, which models the lexical knowledge of English domain. Synonym sets (synsets) are formed by grouping nouns, adjectives, verbs and adverbs. A word can be found in multiple synsets. Each synset essentially represents a concept in the English language. A synset includes the definition of the concept along with an example sentence. Synsets are connected to each other through various relationships such as hypernym, hyponym, synonym, antonym, meronym, holonym.

WordNet [16] contains more than 118,000 different word forms and more than 90,000 different word senses, or more than 166,000 word forms and sense pairs.

WordNet.Net library is the .Net Framework library for WordNet. WordNet.Net was originally created by Malcolm Crowe which is now superceded by several WordNet database versions and library enhancements or bug fixes.

C. Semantic Similarity Measure

Semantic Similarity relates to computing the conceptual similarity between terms which are not necessarily lexically similar. Semantic similarity is measured between two English words by exploring the utility of WordNet, which is a valuable English knowledge base.

Nouns and verbs are arranged into taxonomies, so that concepts and word senses are connected among themselves through various relationships. A semantic relation is one which connects two synsets, whereas a lexical relation is one which connects two members of different synsets. For example, hyponym and hypernym are semantic relations, whereas antonym is a lexical relation.

Although WordNet also contains other Part Of Speech (POS) items, such as adverbs and adjectives, those items are not organized in IS-A hierarchies.

For computing WordNet-based semantic similarity, we employed the open source made available by Thanh Ngoc Dao and Troy Simpson [17]. This incorporated both WordNet.Net library for WordNet 2.1 and the source code that measures semantic similarity. The source code is connected to WordNet 2.1 via WordNet.Net.

Hyponym/Hypernym (or IS-A relations) used to measure the semantic similarity between two synsets and WordNet has 82% of IS-A relations. Parts of speech of "noun-noun" and "verb-verb" are only considered, due to the limitation of IS-A hierarchies.

Semantic similarity is measured between two words for each of its verb-verb/ noun-noun relations. When two words have both noun and verb relations and also semantic similarity exist for both relations then the relation that has higher semantic similarity is considered.

Thanh Ngoc Dao and Troy Simpson [17] addressed various issues for computing semantic similarity.

D. Measuring Similarity

Thanh Ngoc Dao and Troy Simpson implemented the Wu and Palmer method [18] for computing Semantic Similarity between two words or concepts. Wu and Palmer method is an edge based approach which is easy to implement and provides the acceptable accuracy in very simple taxonomy. The measure takes into account both path length and depth of the least common superconcept. Path length is measured in nodes/vertices rather than in links/edges. The length of the path between two members of the same synsets is 1(synonym relations).

Measure of semantic similarity is numbered between 0 and 1. Where 0 signifies little-to-none and 1 signifies extremely high similarity.

Thanh Ngoc Dao and Troy Simpson [17], implemented the equation proposed by Wu and Palmer.

$$\text{Sim}(s, t) = 2 * \text{depth}(\text{LCS}) / (\text{depth}(s) + \text{depth}(t))$$

- Where s and t: denote the source and target words being compared.
- Depth(s): is the shortest distance from root node to a node S on the taxonomy where the synset of S lies.
- LCS: denotes the least common superconcept of s and t.

Examples: The few examples of semantic similarity measure between some sources and targets by using the above Wu and Palmer equation are given below.

Sim(dean, head)= ?
depth(dean) = 9
depth(head) = 6
depth(LCS(person, individual,..)) = 6
Sim(dean, head)= 2*6 / 9+6 = 0.8

Sim(professor, lecturer)= ?
depth(professor) = 9
depth(lecturer) = 8
depth(LCS(educator, pedagogy)) = 7
Sim(professor, lecturer) = 2 * 7/ 9 + 8 =14 / 17 = 0.82

Sim(teacher, instructor)= ?
depth(teacher) = 8
epth(instructor) = 8
depth(LCS(object) = 8
Sim(teacher, instructor) = 2*8 / 8+8 = 1

Intuitively, we got semantic similarity results between two concepts or words based on path length and least common superconcept between them. Semantic similarity results are extremely high i.e 1, if there is no path length between concepts such as teacher and instructor. Semantic similarity results become lower as the distance between concepts and its least common superconcept is increased.

3. EVALUATION

In order to evaluate accuracy of the strategy for ill-defined problems, three English passages which contain 25 ill-defined problems were solved by 10 students of intermediate level. The total 250 solutions of ill-defined problems were collected by students in which 100 student solutions were not considered for evaluation because that student's solutions were exactly matched with the particular solutions of given problems. The other 150 student solutions were considered as a sample size because that were not exactly matched with the particular solutions of given problems. The 150 student solutions were then evaluated by 3 English experts of Isra University and 1 English expert of Degree College Hyderabad. Then the semantic similarity measure technique was run against student solutions and semantic similarity measure values were recorded.

In order to collect student plausible solutions for evaluating the strategy such type of three passages were given to students, where they were asked to answer the cloze passages. The 4 English experts were given a questionnaire of all passages containing all 150 student solutions with particular solutions of given problems along with ranking categories. The experts were asked to rate the acceptability of each student solution on a scale of 1-5, where 1 implied unacceptable, 2 implied not quite acceptable, 3 implied Neutral, 4 implied close to acceptable and 5 implied acceptable.

The collected student's solutions were evaluated twice to validate its plausibleness because the last 4th expert correlation with others was a little lower. The experts ranked students different solutions according to their accuracy and relevance. For the first time evaluation, all 4 experts ranking results are evaluated. For the second time evaluation, first 3 experts ranking results are evaluated because of their more satisfactory correlation that indicate a valid hypothesis and excluding the 4th expert ranking result because of its little lower correlation with others.

The 4 experts agreed among themselves with a correlation of 0.64 with $p<0.05$. The student's solutions were also measured by the semantic similarity approach. The correlation between semantic similarity values and average 4 experts results for each student's solutions is 0.54 with $p<0.05$. Whereas the 3 experts agreed among themselves with a correlation of 0.7 with $p<0.05$ and correlation between first 3 experts and semantic similarity values is 0.58 with $p<0.05$.

A. First Evaluation ROC

For the first time evaluation, all 4 experts ranking results and semantic similarity measure values were evaluated using ROC analysis. The entire student's solutions that had an average of expert ranking above and equal to 4 were considered acceptable or plausible. Set of different cutoff for semantic similarity measure and those students solutions, which scored above and equal to the cutoff were considered acceptable or plausible. Based on average experts ranking and semantic similarity values, evaluated all 150 students solutions using ROC analysis. If the outcome from a semantic similarity measure is positive (above or equal the cutoff) and the human expert ranking result is also positive (above or equal 4), then it is called a true positive (TP); however if the expert ranking result average is negative (less than 4) then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the semantic similarity outcome and the expert outcome are negative, and false negative (FN) is when the semantic similarity measure outcome is negative while the expert ranking is positive. The area under the curve comes out to be 0.71 (Figure 1)

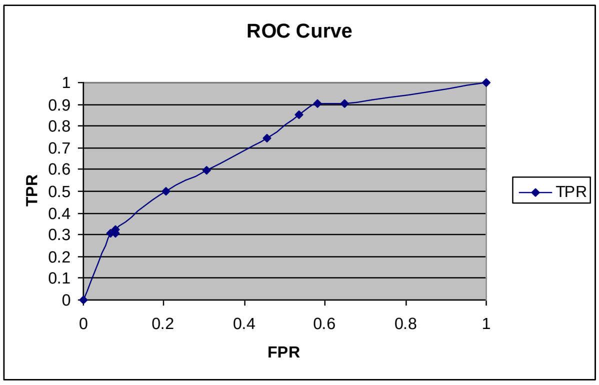

Figure 1. ROC Curve of First Evaluation

B. Second Evaluation ROC

For the second time evaluation, the first 3 experts ranking results excluding the last 4th expert are evaluated due to its lowest correlation. The first 3 experts ranking results and semantic similarity measure values are again evaluated using ROC analysis and the area under the curve of ROC is 0.76 (Fig. 2).

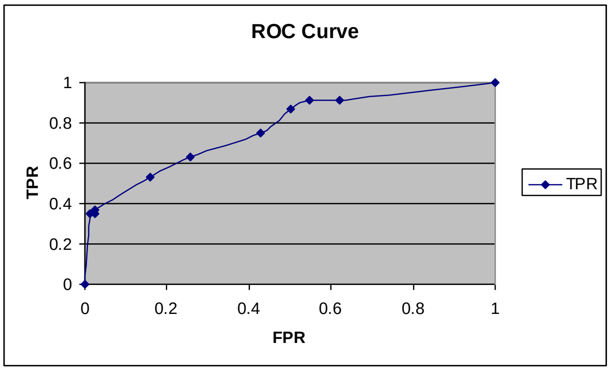

Figure 2. ROC Curve of Second Evaluation

4. DISCUSSION

From the evaluation, it is seen that the last 4th expert has lower correlation that's why the student solutions were evaluated twice, for the first time evaluated by all 4 experts and second time by the first 3 experts.

Those student solutions which are synonyms of the particular solutions have got higher ranking by experts which is "Acceptable-5" and "close to acceptable-4". The strategy also has extremely higher semantic similarity result for that student solutions i-e 1, because both the student solutions and particular solutions of given problems lie on the same synset of 'WordNet ontology'. Hence, it is concluded that the strategy "semantic similarity measure" can result very well to that plausible student solutions which are synonyms of particular solutions.

Whereas the semantic similarity measure responses were not extremely high for those student's solutions which are hypernym/ hyponym of particular solution but according to context it can fit on a blank and also ranked 5 or 4 by expert. Because those students plausible solutions and particular solution don't lie on the same synset and plausible student solutions were on some distance to the particular solution in WordNet ontology.

In the first evaluation, it is seen that 6 FP student solutions were at cutoff 0.96 and in second evaluation that was only 1. The TP in second evaluation was also increased by accepting that 5 FP student solutions at cutoff 0.96 of first evaluation, because that 5 student solutions were ranked 5 or 4 by first 3 experts but not ranked 5 or 4 by the 4th expert. That's why second evaluation resulted in increasing TPR and decreasing FPR at all cutoffs compared to the first evaluation. The correlation of first 3 experts ranking results and experts with semantic similarity measure values were also increased. The area under the curve of ROC is also increased (Figure 2).

5. CONCLUSIONS

The ROC curve shows fair results for our method of accepting plausible solutions; however, the results could be improved. Other techniques for accepting plausible solutions should also be investigated that could possibly lead to improved results.

This paper addresses the problem of accepting plausible student solutions in English ill-defined domain but by using our proposed strategy, ill-defined problems of other ill-defined domains could also solve this problem.

This paper is focused only on assessment of ill-defined problems in English domain that could extend to develop ITS prototype which provides feedback to students. Assessment could be extended to accept also those plausible student solutions which are acceptable according to context. More evaluation of learning outcome for ITS could also be conducted.

Further English tutoring system could be extended to assess best solution among multiple acceptable solutions. As ill-defined problem possess multiple solutions and contain uncertainty about which solution is best. To select best solution among multiple solutions, assess the viability of alternative solutions by constructing arguments and articulating personal beliefs. The argumentation can provide a valuable assessment of the learner's problem solving ability.

6. REFERENCES

[1] Lynch, C.F, Ashley, K.D., and 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.

[2] Aleven, V., Ashley, K., Lynch, C. and Pinkwart, N. (2008), Preface, Intelligent Tutoring Systems for Ill-Defined Domains: Assessment and Feedback in Ill-Defined Domains, the 9th Conference on ITS.

[3] Matsuda, N., and VanLehn, K. (2005) Advanced Geometry Tutor: An intelligent tutor that teaches proof-writing with construction. In C.-K. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, pp. 443-450.

[4] Melis, E., and Siekmann, J. (2004) ActiveMath: An Intelligent Tutoring System for Mathematics. *Seventh International Conference Artificial Intelligence and Soft Computing (ICAISC).* In L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L.A. Zadeh (Eds.), Lecture Notes in Artificial Intelligence, Springer-Verlag, Vol. 3070, pp. 91-101.

[5] Corbett,A.T., Koedinger.K.R and Anderson, J.R (1997) *Intelligent Tutoring Systems*, Handbook of Human-Computer Interaction, Second Completely Revised Edition in M. Helander, T. K. Landauer, P. Prabhu (Eds), Elsevier Science B. V., Chapter 37.

[6] Mitrovic, A. (1997) *SQL-Tutor: A preliminary report.* Technical report TRCOSC 08/97, Computer science department, University of Canterbury.

[7] Kazi, H., Haddawy, P., Suebnukarn, S. Expanding the Space of Plausible Solutions in a Medical Tutoring System for Problem Based Learning. *International Journal of Artificial Intelligence in Education 19*, 3 (2009), 309-334.

[8] Fum, D., Giangrandi, O. and Tasso, C. (1992) The Use of Explanation-Based Learning for Modeling Student Behavior in Foreign Language Tutoring. In M. L. Swartz, M. Yazdani (Eds.) *Intelligent Tutoring Systems for Foreign Language Learning,* Berlin: Springer Verlag, pp.151-170.

[9] Schuster, E. (1986) The role of native grammars in correcting errors in second language learning, *Computational Intelligence,* 2, 9398.

[10] Bos, E. and van de Plassche, J. (1994) A Knowledge-Based, English Verb-Form Tutor. *Journal of Artificial Intelligence in Education,* Spengels, Vol. 5, No. 1, pp.107-129.

[11] Boucher, P. and Danna, F. et Pascale Sebillot (1993) Compounds: an intelligent tutoring system for Learning to Use Compounds in English. *Computer Assisted Language Learning (CALL),* ISSN 1166-8687, Vol.6, No. 3, pp.249-272.

[12] Mayo, M., Mitrovic, A. and McKenzie, J. (2000) CAPIT: An intelligent tutoring system for Capitalisation and Punctuation. *Proceedings of the International Workshop on Advanced Learning Technologies,* pp. 151-154.

[13] Virvou, M., Maras, D., and Tsiriga, V. (2000) Student Modelling in an Intelligent Tutoring System for the Passive Voice of English Language. *In EDUCATIONAL TECHNOLOGY & SOCIETY, Journal of International Forum of Educational Technology & Society and IEEE Learning Technology Task Force*., Vol.4, No, 3, pp.139-150.

[14] Collins-Thompson, K. and Callan, J. (2004) Information retrieval for language tutoring: An overview of the REAP project. In Proceedings of the Twenty Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK.

[15] Brown, J. and Eskenazi, M. (2004) Retrieval of authentic documents for reader-specific lexical practice. *In Proceedings of InSTIL/ICALL Symposium*, Venice, Italy.

[16] Miller, G.A. (1995) WordNet: A Lexical Database for English. *Communication of ACM,* Vol. 38, No. 11, pp. 39-41.

[17] Troy Simpson and Thanh Dao, (2005, October 01), "WordNet-based semantic similarity measurement", (The Code Project) Available: http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx

[18] Wu, Z and Palmer, M. (1994) Verb Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meeting of the Associations for computational Linguistics (ACL'94), pp. 133-138.

[19] "READING #1", (2003, October 07), (INTERLINK Language Centers), Available: http://eslus.com/LESSONS/READING/CLOZE/R1.HTM