Using an Implicit Method for Coreference Resolution and Ellipsis Handling in Automatic Student Answer Assessment

Rajendra Banjade, Vasile Rus, Nobal B. Niraula

Department of Computer Science
Institute for Intelligent Systems
The University of Memphis
{rbanjade, vrus, nbnraula}@memphis.edu

Abstract

The automatic student answer assessment problem is challenging because it requires natural language understanding. This problem is even more challenging in conversational Intelligent Tutoring Systems (ITS) because in such conversations the speakers develop common ground as the dialogue proceeds, which means contextual information from previous utterances in the dialogue is heavily relied upon to understand a speaker's utterances. Different linguistic phenomena should be addressed in order to improve the performance of automatic answer assessment systems in conversational ITS. Two such important phenomena are: *references to entities mentioned earlier in the dialogue* and *ellipsis (i.e., answers with contextually implied parts)*. In this paper, we present an implicit approach to resolving coreferences and handling elliptical responses in the context of automatic student answer evaluation in dialogue based intelligent tutoring systems.

Introduction

Research in educational technologies has enabled moving towards building artificial agents, such as dialogue based intelligent tutoring systems (ITSs; Rus, D'Mello, Hu, & Graesser, 2013; Evens & Michael, 2006; Graesser et al., 2004), that offer tailored instruction to each individual learner in order to maximize students' learning and ultimately mastery of the target domain.

Automatic evaluation of student answers is one of the critical components of dialogue based intelligent tutoring systems because accurate assessment of student input is needed in order to provide effective feedback, which in turn impacts learning. In conversational ITSs, students provide their answers in natural language and the computer tutor has to evaluate these answers (typically by comparing them with expert answers) before providing a response. The evaluation of student responses, which are open ended

in nature (i.e., the student may utter whatever they wish), is an extremely challenging problem (Banjade, 2014; Nielsen, 2009). From a technical standpoint, it is difficult mainly because it requires natural language understanding, which is a challenging task because it requires world and domain knowledge, which are two intractable problems in Artificial Intelligence. A quick survey of published works on student answer assessment has shown that the ideas of semantic similarity and textual entailment have been borrowed directly such that a student answer is compared with an expected answer by simply extending word-to-word semantic similarity measures ignoring important aspects of natural language processing that must be accounted for (Graesser et al., 2004; Mohler, M., & Mihalcea, 2009; Rus & Lintean, 2012; Dzikovska et al., 2013; Murrugarra et al., 2013). Indeed, the (near) perfect evaluation of student responses requires many additional processing steps, for example, coreference resolution, ellipsis handling, negation handling etc. In this paper, we present an implicit approach to handling two important linguistic phenomena present in student responses in dialogue based intelligent tutoring systems: coreference *resolution* and *ellipsis handling*.

Coreference occurs when two or more expressions in a text refer to the same entity (person, thing, or noun phrase); we say that they have the same referent. For example, students can write "*they*" to refer to a set of forces mentioned in a problem description during problem solving activities in tutoring. Indeed, this is a commonly occurring phenomenon in computer-based tutoring. For example, Niraula and Rus (2014) found 5,881 pronouns in 25,945 student turns which accounts to one pronoun every four utterances on average. They looked at anaphoric references only, i.e. the use of pronouns to refer to other entities and ignored the use of non-anaphoric phrases to refer to entities. Students also use co-referring noun phrases, such as *these forces*. Such referring phrases need to be resolved to properly assess the correctness of student

responses. For instance, coreferent such as *these forces* should be replaced with the actual forces these phrase refers to before attempting to assess a student answer containing these coreferring phrase.

The other linguistic phenomena that we address in this paper are about incomplete (or elliptical) utterances. Elliptical student responses are common in conversations between humans even when they are instructed to produce more syntactically and semantically complete utterances (Carbonell, J. G., 1983).

Elliptical utterances range from syntactically incorrect sentence fragments to sentences that fail to include all requisite semantic information. Such elliptical utterances are common in tutor-student interactions in ITSs. Though tutoring systems are so designed to generate semantically and syntactically complete utterances, the student utterances are often elliptical. In many cases student utterances cannot be understood in isolation; they only make sense when interpreted within context. In general, handling elliptical utterances is a very difficult problem for natural language understanding systems. In our case, we propose an indirect approach to handling elliptical utterances together with solving coreference resolution.

The proposed approach was evaluated on a dataset (called DeepEval; described in the Data section) from 618 student responses collected from an experiment with DeepTutor (Rus, D'Mello, Hu, & Graesser, 2013).

Student Responses Analysis

We illustrate in this section the complexity of student answer assessment with an emphasis on exemplifying the importance of reference resolution and ellipsis handling. Student responses to the same tutor question can vary a lot and ITSs must be able to handle and correctly assess all types of responses. We present below an example of a Physics problem and a tutor hint in the form of a question. Actual student responses (or part of their responses) for the question are presented in Table 1. It should be noted that the list of student response is by no means exhaustive. We also show a list of expert-generated answers that represent the current logical step in the ideal solution to the problem.

Example 1: **Problem Description:** *To rescue a child who has fallen down a well, rescue workers fasten him to a rope, the other end of which is then reeled in by a machine. The rope pulls the child straight upward at steady speed.*
Question: *How does the amount of tension in the rope compare to the downward force of gravity acting on the child?*

Reference Answers:

✓ *The amount of tension in the rope is the same as (equal to) the magnitude of the downward force of gravity.*
✓ *Since the child is being raised straight upward at a constant speed, the net force on the child is zero and all the forces balance. That means that the tension in the rope balances the downward force of gravity.*

ID	Student response
1	*these forces balance each other*
2	*The tension is equal to the force of gravity*
3	*They are equal.*
4	*Equal*
5	*the tension force is balanced by the weight of the person*
6	*they are equal and opposite in direction*

Table 1: Some of the student answers (or part of their answers) for the question given in Example 1.

The student answer (3) in Table 1 is using "*they*" in his/her response which, in this example, is not referring to a group of people but to *the amount of tension in the rope and the downward force of gravity*. Similarly, "*these forces*" in answer (1), same as "*they*" in answer (3), is referring to the *tension force and the force of gravity*. The student answer (4) in Table 1 is missing important details. It does not even use co-referring expressions to refer to the important information needed to understand this answer. However, the missing information is inferred from the context (i.e., the tutor question in this case). A human teacher most probably would give credit to this student answer as the student mentioned the most important piece of information, i.e. equal, in the reference answer. One might suggest that the problem could be avoided in human-machine communication by training human users to employ only syntactically and semantically complete utterances. However, empirical research shows that this is not feasible (Carbonell J. G., 1983). Constraining man-machine communication to certain types of utterances and restrict the use of utterances normally employed by humans would could have negative consequences in terms of user perception of the tutoring system and even the effectiveness of the tutoring. Ideally, the learner-system interaction should be as natural as possible, without any constraints. That is, students should be allowed to respond and articulate answers as naturally as possible even if they use coreferring expressions and ellipsis.

Related Works

Automatic Answer assessment. Assessing students' natural language answers has been explored by many research groups. The previous major efforts were spent on two tasks: automated essay scoring, e.g. SAT essay grading,

(Landauer et al., 1997; Wiemer-Hastings & Graesser, 2000; Lemaire & Dessus, 2001; Leacock, 2004) and short answer evaluation (Graesser et al., 2004; Nielsen, 2009; Rus, Niraula, & Banjade, 2013, Mohler & Mihalcea, 2009). The short-answer evaluation is the focus of our work. As mentioned, in ITSs students' responses are automatically compared to expert answers through a semantic similarity step. When student answers are compared to the expert answers, we argue that ellipsis and co-referents must be handled such that any contextual information is fully available during the semantic similarity operation between the student response and the expert answer.

Coreference Resolution. The coreference resolution systems proposed over the past several decades are both supervised methods (Bengtson, E., & Roth, 2008) and unsupervised methods (Raghunathan et al. (2010). However, as discussed by Stoyanov et al. (2010), we have little understanding of which aspects of the coreference resolution problem are handled well or poorly by state-of-the-art systems. Also, most of the systems work on written texts but our work is on dialogue, i.e. conversational texts. A related work by Nobal and Rus (2014) is on using machine learning approach for pronominal anaphora resolution in dialogue based ITS. In our case, we implicitly cover different forms of coreferences including pronouns, coreferring noun phrases etc.

Ellipsis Resolution. Ellipsis resolution is the task of recovering the implied antecedent information. Elliptical utterances range from syntactically incorrect sentence fragments to sentences that fail to include all requisite semantic information. Most of the previous work focuses on reconstruction of the elided fragment of text where the ellipsis is present and is related to a previous text fragment within the same sentence or utterance (Giuliani et al., 2014; Illaraza-Sanchez, Hontoria, & Verdejo Maillo, 1990; Kehler, 1993; Lapin & Shih, 1996). They mostly exploit syntactic information to recover the missing information. However, in dialogue based systems, semantically incomplete utterances should be interpreted based on the contextual information found in previous utterances (i.e., dialogue history) and other related information. For instance, Schlangen (2003) resolved intended content of non-sentential utterances as a by-product of the establishment of coherence in dialogue. Fukumoto et al. (2004) proposed an approach where follow-up inquiry questions with elliptical construct are completed based on previous questions of the same speaker. In our case, student ellipsis is more likely linked to previous tutor hints in the form of questions and therefore we address across-speaker ellipsis and the type of texts are question-answer pairs. There is no research on ellipsis handling in automated answer assessment to the best of our knowledge.

Implicit Coreference Resolution and Ellipsis Handling

The proposed solution for the two problems - *coreference resolution* and *ellipsis handling* - is based on the idea of using information from the previous dialogue context in order to make the student responses self-contained. However, we do so without explicitly addressing the two problems of coreference resolution and ellipsis handling which is difficult; furthermore, state-of-the-art solutions to these problems are error prone because the two problems are very challenging. Therefore, instead of directly resolving coreferences and ellipsis we simply align concepts from the dialogue context to the student answer.

This idea fits well with the basic principles (steps) applied in evaluating student answers such as checking whether the target concepts in the reference answer are present in the student response. We us principal chunks, i.e. phrases, as concepts (as described by Stefanescu et al., 2014). We extract concepts from the student response, tutor question, and the reference answers. After that, the concepts in the three texts are optimally aligned (optimal alignment proposed by Rus and Lintean, 2012) with each other (as illustrated in Figure 1). Then, the following rules are applied to assess the student response which in a way indirectly resolve coreferences and ellipsis.

1. Do not take into account a concept if that concept in the student response matches (aligns) with a concept present in the tutor question, i.e., if it is clearly implied by the context provided by the previous tutor question.
2. Do not give any credit for concepts in the reference answer if they are present in the question itself irrespective of whether the concept is present in the student answer. That is, if the tutor mentions some concepts in the question if the student repeats them in her student response, no credit is given for the concept.
3. Only the remaining concepts in student response and reference answer are used to calculate the similarity score (the extent to which the reference answer is covered by the student answer).

How does it address ellipsis handling and coreference resolution?

As briefly mentioned earlier, our basic idea is to align concepts in the student answer, the dialogue context, and the reference answers. This strategy effectively means that the concepts in the tutor question which the student may explicitly, indirectly (through a pronoun or other co-referring expressions), or implicitly (ellipsis) mention are assumed to be part of the student answer with the caveat that the student would not be given credit for concept he

does not explicitly articulate; however, the advantage is that the student answer will be properly evaluated by implicitly handling implied concepts as explained next.

A first question to address is how much dialogue context should be considered for the proposed method. Since most of the time students co-refer, if at all, to entities or noun phrases within the answers themselves or the most recent utterances in the dialogue, we have considered only last tutor utterance (i.e., question) as context. Indeed, Niraula et al. (2014) found that about 84.76% of the pronouns refer to an entity in the answer text itself or the last tutor utterance (53.22% in last tutor utterance and 31.54% within the answer itself). Therefore, when students write answers containing pronouns or other referring expressions the referred concepts are most likely mentioned earlier in the same student response or most recent tutor question. If the implied entity is in the student answer itself, by aligning the student answer with the reference answer the referred entity will be aligned with an entity in the reference answers and by default the co-referring expression will be indirectly aligned. If the referring expression in the student answer refers to an entity in the previous tutor question and which is typically also found in the reference answer as

well (that is how the tutor hints are authored by experts), the entity in the reference answer would align with the entity in the question and therefore indirectly with the referring expression in the student answer.

Similarly, if a concept present in the question is also present in the reference answer but it is missing in the student answer, e.g. due to an ellipsis in the student response, then by aligning the concept in the question with the concept in reference answer would effectively work as aligning the implied concept in the student answer with the concept in reference answer. Thus, it practically resolves the ellipsis and makes the student utterance look like being more complete without explicitly making it so. As an example we show in Figure 1, a student answer in which the student did not mention explicitly *"when the rocket stops pushing."* This elided part is present in the reference answer as well as in the question. Students see the question and sometimes respond elliptically omitting this fragment as it is implied by the question. By aligning the concepts – *rocket, stops,* and *pushing* - present in question and reference answer, it effectively works as if the student has explicitly mentioned these concepts.

Problem description: A rocket is pushing a meteor with constant force. At one moment the rocket runs out of fuel and stops pushing the meteor. Assume that the meteor is far enough away from the sun and the planets to neglect gravity.

Question: How will the meteor move after the rocket stops pushing?

Student response: *it will move at a constant speed*

Reference answer: When the rocket stops pushing, no forces are acting on the meteor anymore and therefore will move with constant velocity in a straight line

Concepts in student answer:
a1. [VP moves/move/VBZ] – q2, r7
a2. [NP constant/constant/JJ speed/speed/NN] – r8

Concepts in question:
q1. [NP meteor/meteor/NN] – r6
q2. [VP move/move/VB] – r7
q3. [NP rocket/rocket/NN] – r1
q4. [VP stops/stop/VBZ] – r2
q5. [VP pushing/push/VBG] – r3

Concepts in Reference answer:
r1. [NP rocket/rocket/NN] – q3
r2. [VP stops/stop/VBZ] – q4
r3. [VP pushing/push/VBG] – q5
r4. [NP zero/zero/CD forces/force/NNS]
r5. [VP acting/act/VBG]
r6. [NP meteor/meteor/NN] – q1
r7. [VP move/move/VB] - a1, q2
r8. [NP constant/constant/JJ velocity/velocity/NN] – a2
r9. [NP straight/straight/JJ line/line/NN]

Figure 1: Image showing the alignments of concepts among student answer, question, and a reference answer.

Data

In order to evaluate the effectiveness of the proposed approach, we collected and annotated (binary judgment) a dataset (called DeepEval) that was extracted from anonymized records of tutor-student conversations in one experiment with DeepTutor, a state-of-the-art intelligent tutoring system. The experiment involved 41 college students. Each student was given 9 conceptual physics problems to solve during a one-hour tutoring session. The interaction between the learner and the system was natural language dialogue using typed text, i.e. chatroom-like conversations. During the tutorial dialogue, the intelligent tutoring system automatically assessed the correctness of student responses by comparing the student responses with the reference answers provided by domain experts. We randomly selected a subset of the dialogue interactions and manually annotated student answers for correctness using a binary judgment (correct/incorrect). That is, the dataset contains naturally occurring texts. While students were encouraged at the beginning of the interaction with DeepTutor to write complete sentences as much as

possible, they were free to type anything and anyhow they wanted. Table 2 presents the summary of the dataset.

Parameter	Value
Number of task solving dialogues	198
Total number of instances	618
Total number of positive instances	358
Avg. number of words in problem description	25.96
Avg. number of words in questions	15.77
Avg. number of words in student answers	14.93
Avg. number of words in reference answers	17.07
Annotation agreement (Kappa, %)	83 (92%)

Table 2: DeepEval dataset summary

Instances were annotated by two annotators and their agreement was calculated (Kappa=0.83 and percentage of agreement 92%).

Experiments and Results

Experiments

First, we preprocessed student responses. Spelling corrections were handled using edit distance based method implemented in Jazzy API[1]. Lemmatization and Parts-Of-Speech tagging were performed using Stanford CoreNLP[2] package. Then, we extracted concepts from student answer, reference answer(s), and tutor question using Stanford CoreNLP package as described by Stefanescu, Banjade, and Rus (2014). To evaluate the correctness of student responses, we developed a scoring model with four different features:

Expectation coverage score (ECC): It quantifies how much of the reference answer is covered by the student answer. If there are multiple reference answers, the most covered reference answer is taken. To align concepts, words within concepts were aligned optimally (modeled as a well-known combinatorial optimization problem; Rus & Lintean, 2012). For this experiment, we used strict measure for word alignment based on the synonymy relation in WordNet (Fellbaum, C., 1998). If words (lemmatized) are same or hold synonym relation in WordNet, they are perfectly similar. Once concept-to-concept similarity scores are calculated, concepts are finally aligned using optimal alignment method similar to word-to-word alignment method used for concept-to-concept similarity calculation.

Presence of contradicting concept (PCC): The presence of contradicting concepts (e.g., greater and smaller) or disjoint set of concepts (e.g., equal and greater; they do not hold antonymy relations but they are different) in the

student answer is a sign that the student answer contradicts (parts of) the expected answer. In this experiment, we only looked at antonym relation in WordNet to check whether concepts are contradicting.

Uncovered concepts in the reference answer (UCRA):
It is the number of uncovered concepts in the reference answer normalized by the total number of concepts in it.

Uncovered concepts in the student response (UCSR):
It is the number of uncovered concepts in the student response normalized by the number of concepts present in the response itself.

Based on the above features, we trained a logistic model and evaluated it using 10-fold cross validation (using Weka tool). The performance was measured in terms of accuracy (Acc), precision (Prec), recall (Rec), F1-score (F1), and reliability (Kappa). The performance of the method with and without implicitly resolving coreferences and elliptical responses was compared and the results are presented next.

Results

Table 3 presents the comparative results with and without implicit coreference resolution and ellipsis handling (represented as ICE and NoICE respectively). We grouped model features as F1 (all four features), F2 (ECC and PCC), and F3 (ECC only).

Run	Acc	Prec	Rec	F1	Kappa
NoICE_F1	66.34	66.00	66.30	64.80	27.61
NoICE_F2	65.04	64.40	65.00	63.90	25.41
NoICE_F3	61.97	61.20	62.00	61.10	19.61
ICE_F1	70.87	70.60	70.90	70.50	39.07
ICE_F2	**70.87**	**70.60**	**70.09**	**70.60**	**39.27**
ICE_F3	69.25	68.90	69.30	68.70	35.40

Table 3: Scoring results obtained with and without implicit coreference resolution and ellipsis handling.

The results show that after handling the coreference resolution and elliptical responses, the performance is improved significantly. The F1-score went up from 64.80 (kappa 27.61) to 70.60 (kappa 39.27). The best results were obtained by using only two features - ECC and PCC. The features that measure uncovered part of reference answer and uncovered part of student answer did not significantly improve the results.

Discussions and Conclusions

It is human nature that we tend to produce syntactically simpler and shorter utterances despite any training given to produce complete utterances (empirically verified by Carbonell's study, 1983). This is particularly true for conversational contexts. Similarly, students were advised to write as complete as possible during DeepTutor

---

[1] http://jazzy.sourceforge.net/
[2] http://nlp.stanford.edu/software/corenlp.shtml

experiment. However, the reality is that many of them did not follow these instructions. Forcing students to write responses with semantically complete or self-contained responses is undesirable because they might be distracted and demotivated using intelligent conversational systems. Ideally, students should be allowed to express themselves as naturally as possible. Automatic handling of any student response is thus needed. If they can freely express their responses and the computer tutor understands them, it will actually reduce the gap between human-to-human tutoring and machine-to-human tutoring, which is the ultimate goal of artificial intelligence research in conversational intelligent tutoring systems. This work is an important step towards facilitating human-level quality of conversation with intelligent tutoring systems.

To conclude, we obtained significant improvement in automatic student answers evaluation in dialogue based intelligent tutoring systems by using indirect approach for addressing two important linguistic phenomena present in student responses: *coreferences* and *elliptical responses*. In the future, we intend to look at dialogue history beyond tutor's last utterance and the problem description as additional context and to better understand the relationship between dialogue context, problem context and the production and resolution of elliptical responses and coreferences.

Acknowledgments

References

Bengtson, E., & Roth, D. (2008, October). Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 294-303). Association for Computational Linguistics.

Carberry, S. (1989). A pragmatics-based approach to ellipsis resolution. *Computational Linguistics*, *15*(2), 75-96.

Carbonell, J. G. (1983, June). Discourse pragmatics and ellipsis resolution in task-oriented natural language interfaces. In *Proceedings of the 21st annual meeting on Association for Computational Linguistics* (pp. 164-168). Association for Computational Linguistics.

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., ... & Dang, H. T. (2013). *SemEval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge*.

Evens M., & Michael J. (2006), *One-on-One Tutoring by Humans and Computers*. Lawrence Erlbaum Associates.

Fellbaum, C. (1998). *WordNet*. Blackwell Publishing Ltd.

Fukumoto, J., Niwa, T., Itoigawa, M., & Matsuda, M. (2004, June). RitsQA: List answer detection and context task with ellipses handling. In *Working notes of the Fourth NTCIR Workshop Meeting, Tokyo, Japan* (pp. 310-314).

Giuliani, M., Marschall, T., & Isard, A. (2014, June). Using Ellipsis Detection and Word Similarity for Transformation of Spoken Language into Grammatically Valid Sentences. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (p. 243).

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. Behavior Research Methods, Instruments, & Computers, 36(2), 180-192.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389-405

Mohler, M., & Mihalcea, R. (2009, March). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 567-575). Association for Computational Linguistics.

Murrugarra, N., Lu, S., & Li, M. L. (2013). Automatic Grading Student Answers.

Nielsen, R. D., Ward, W., & Martin, J. H. (2009). Recognizing entailment in intelligent tutoring systems. *Natural Language Engineering*, *15*(04), 479-501.

Niraula, N. B., & Rus, V. (2014). A Machine Learning Approach to Pronominal Anaphora Resolution in Dialogue Based Intelligent Tutoring Systems. In *Computational Linguistics and Intelligent Text Processing* (pp. 307-318). Springer Berlin Heidelberg.

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., & Manning, C. (2010, October). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 492-501). Association for Computational Linguistics.

Rus, V., D'Mello, S. K., Hu, X., & Graesser, A. C. (2013). Recent Advances in Conversational Intelligent Tutoring Systems. AI magazine, 34(3), 42-54.

Rus, V., & Lintean, M. (2012, June). A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 157-162). Association for Computational Linguistics.

Rus, V., Niraula, N., & Banjade, R. (2013). Similarity measures based on latent dirichlet allocation. In *Computational Linguistics and Intelligent Text Processing*(pp. 459-470). Springer Berlin Heidelberg.

Ştefănescu, D., Banjade, R., & Rus, V. (2014). A Sentence Similarity Method Based on Chunking and Information Content. In *Computational Linguistics and Intelligent Text Processing* (pp. 442-453). Springer Berlin Heidelberg.

Stoyanov, V., Gilbert, N., Cardie, C., & Riloff, E. (2009, August). Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2* (pp. 656-664). Association for Computational Linguistics.