The DARE Corpus: A Resource for Anaphora Resolution in Dialogue Based Intelligent Tutoring Systems

Nobal B. Niraula, Vasile Rus, Rajendra Banjade, Dan Stefanescu, William Baggett, Brent Morgan
Department of Computer Science, Department of Psychology, Institute for Intelligent Systems
The University of Memphis
Memphis, TN 38152
E-mail: {nbnraula,vrus,rbanjade,dstfnscu}@memphis.edu, {writebill, brent.morgan}@gmail.com

Abstract

We describe the DARE corpus, an annotated data set focusing on pronoun resolution in tutorial dialogue. Although data sets for general purpose anaphora resolution exist, they are not suitable for dialogue based Intelligent Tutoring Systems. To the best of our knowledge, no data set is currently available for pronoun resolution in dialogue based intelligent tutoring systems. The described DARE corpus consists of 1,000 annotated pronoun instances collected from conversations between high-school students and the intelligent tutoring system DeepTutor. The data set is publicly available.

Keywords: Anaphora Resolution, Dialogue Systems, Tutoring Systems

1. Introduction

We present in this paper the DARE corpus, an annotated data set for fostering the development of anaphora resolution algorithms in dialogue systems. The task of anaphora resolution is to identify the referent of a pronoun in dialogue or discourse. In particular, we focus on the task of anaphora resolution in the context of dialogue-based Intelligent Tutoring Systems (ITSs).

ITSs form a category of advanced educational technologies that tailor instruction to each individual student in order to maximize learning for every student. Indeed, ITSs have already proven to be very effective at inducing learning gains in students (Rus et al., 2013a). In dialogue based ITSs, students typically solve problems and get help, as needed, by having a conversation with the computer tutor. If students struggle, the computer tutor will provide hints in the form of leading questions. The goal is to make the student solve the problem by himself instead of telling the answer. The student will eventually be told the correct solution in case everything else fails.

A critical step in managing the dialogue and the quality of feedback the system provides to the student is assessing the correctness of student natural language input. As part of this step, it is necessary to resolve any pronouns the student utterance might contain.

Consider the real student-tutor interaction below from the intelligent tutoring system DeepTutor(Rus et al., 2013b):

PROBLEM: A mover pushes a desk with constant velocity V_0 across a carpeted floor. Suddenly, the mover stops pushing. What can you say about the motion of the desk after the mover stops pushing ? Explain why.
STUDENT ANSWER: The desk will stop moving because ***it*** *was only moving due to the applied force of the mover pushing on* ***it****.* ***It*** *does not have a constant velocity or acceleration to keep* ***it*** *going.*

The student answer in the example above has four pronouns, all referring to *desk*. In fact, students use pronouns quite frequently in tutorial dialogues. For example, Niraula et al. (2013) reported a total of 5,881 pronouns in 25,945 student turns. As already mentioned, to assess the correctness of student responses with pronouns, the referent(s) of the pronouns must be found. Incorrect assessment of student responses in ITSs could lead to incorrect feedback provided by the system which, in turn, could frustrate students sometimes to the point of quitting. The DARE corpus presented in this paper will provide a much needed resource to develop advanced anaphora resolution algorithms for dialogue-based ITSs.

While data sets for general purpose anaphora resolution have been developed before, they are not suitable for dialogue based tutoring systems. For example, input to general purpose anaphora resolution algorithms is normally a few sentences or paragraph(s) whereas in tutorial dialogues the input would be the whole dialogue history as well as the current instructional task, i.e. Physics problem in our case, which offers the broader context of the dialogue.

To the best of our knowledge, no data set is currently available for pronoun resolution in a dialogue based tutoring context. An annotated data set such as the DARE corpus discussed here would be a valuable resource towards developing highly accurate solutions to the task of anaphora resolution in tutorial dialogue. This in turn will improve the student response assessment methods. In order to build the DARE corpus, we have automatically extracted and then annotated 1,000 unique instances of student use of pronouns from dialogues between high-school students and DeepTutor. We describe in this paper the corpus development process as well as a summary of the resulted corpus.

The paper is organized as follows. In section 2., we present related works. In section 3., we describe creation of DARE corpus and corresponding annotation guideline. Section 4. describes experiments. We conclude the paper in section 5.

2. Related Work

Anaphora resolution is a special case of a broader problem in natural language processing (NLP) called coreference

Table 1: Use of pronouns in students' responses

(a) Intra-turn :
TUTOR:What does Newton's second law say?
STUDENT:for every force, there is another equal force to counteract *it*
(b) Inter-turn immediate:
TUTOR:What can you say about the acceleration of the piano based on Newton's second law and the fact that the force of gravity acts on the piano?
STUDENT: *It* remains constant.
(c) Inter-turn history:
TUTOR: Since the ball's velocity is upward and its acceleration is downward, what is happening to the ball's velocity?
STUDENT: increasing
TUTOR: Can you please elaborate?
STUDENT: *it* is increasing

resolution, which is the task of identifying all co-referents of an entity in texts. Several methods have been proposed for coreference/pronoun resolution that range in applicability from general purpose texts to biomedical texts to multiple languages to spoken dialogues (Poesio and Kabadjov, 2004; Versley et al., 2008; Mitkov et al., 2002; Qiu et al., 2004; Rahman and Ng, 2009; Su et al., 2008a; Stent and Bangalore, 2010). The methods were evaluated against various data sets corresponding to the target domain. For general purpose solutions, the MUC-6 (Message Understanding Conference) data set is used. For multiple languages, the data set from Task 1 of SemEval-2010 is quite popular (Recasens et al., 2010). The CHILD corpus is famous for human spoken dialogue (Bangalore et al., 2008). Similarly, popular data sets in biomedical text analysis are the GENIA corpus (Kim et al., 2003), the GNOME corpus (Poesio, 2004), and the MedCo corpus (Su et al., 2008b).

Although various data sets were developed previously, they are not suitable to all domains because coreference and its special case of anaphora resolution has its peculiarities in different domains and text genres. In the case of tutorial dialogue anaphora no such resource exists, which motivated the development of the DARE corpus.

3. The DARE Corpus

This section presents first the types of anaphora we identified in tutorial dialogues and then describes the process we used for the corpus creation, including the annotation guidelines.

3.1. Anaphora Types in Tutorial Dialogue

We have identified three types of anaphora in student utterances. They include *Intra-turn*, *Inter-turn intermediate* and *Inter-turn history* anaphora - see Table 1. In the case of *Intra-turn* anaphora, the referents are found within the student's current dialogue turn. In *Inter-turn intermediate* anaphora, the referents lie in the most recent tutor turn and in *Inter-turn history* anaphora, the referents are located in earlier dialogue turns or even the problem description.

3.2. Corpus Creation

In order to create the DARE Corpus, we started by extracting 1,000 pronoun instances from student-tutor interaction logs collected during one experiment involving high-school students interacting with the intelligent tutoring system DeepTutor in the domain of conceptual Physics (Rus et al., 2013a). A typical collected instance is presented in Table 2.

INSTANCE: 3624
#FILE: VM_LV03_PR06_dh335-022813.txt
PROBLEM: A stuntman must drop from a helicopter onto a target on the roof of a moving train. The plan is for the helicopter to hover over the train, matching the train's constant speed before the stuntman drops.
Q2:Where should the helicopter be positioned relative to the target? Please begin by briefly answering the above question. After briefly answering the above question, please go on to explain your answer in as much detail as you can.
A2:in front of the target due to wind resistance
Q1:Let me try again. Which principle can be applied when the motion of an object is complex, for instance, it can be thought of as motion in two perpendicular dimensions?
A1:decomposition
Q: What can you say about the motion of the stuntman after he jumps?
A: it will be parabolic

Table 2: A typical instance for anaphora resolution

Each instance has a unique id (e.g. 3,624 in the example) and the name of the log file from which the instance was derived. Current student's response is designated by A(nswer) and the corresponding utterance from the tutor (which in this case is DeepTutor) that triggered the response A, is denoted by Q(uestion), which is the most recent hint from the system which is typically in the form of a question. Previous student answers are denoted with A1, A2, and so on, while previous DeepTutor turns are denoted with Q1, Q2, and so on. The goal here is to resolute pronouns in A to their referents, which could be in the same student response A, the previous tutor turn, earlier in the dialogue history, the common ground built so far by the two conversation partners, or even the description of the current problem that the student is assigned to solve.

Once the set of 1,000 instances was collected from the DeepTutor dialogues, we proceeded with creating the annotation guidelines which borrow some ideas from the guidelines used for annotating the data set used in MUC-6 [1]. However, due to the peculiarities of tutorial dialogues, our guidelines are quite different.

3.2.1. Annotation Guideline

The first task during the annotation is to mark the referent of each candidate pronoun with a special tag and assign each

[1] http://www.cs.nyu.edu/cs/faculty/grishman/muc6.html

referent an unique id. Referents corresponding to anaphoric pronouns are either simple nouns or noun phrases, which is the reason we use <np> tags to surround each referent (we do not consider chains of pronouns). The pronouns themselves are marked with tags. Both the referents and pronouns are assigned unique ids using an *id* attribute associated with the <np> and tags.

Because a noun phrase can be complex, meaning that a noun phrase can be a part of a longer noun phrase, it is necessary to specify how to select the boundaries of the referent. For example, in the sentence: *The red ball which was thrown vertically up strikes the target first, and thus it returns the earliest.*, the possible referents *ball*, *red ball*, and *the red ball* are all part of the longer noun phrase *the red ball which was thrown vertically up*. Which referent should we pick for the pronoun *it*: "ball", "red ball", "the red ball", or "the red ball which was thrown vertically up" ? To address this issue, we chose to mark as the referent the noun phrase that fully and clearly describes the referent, which in this case would be *the red ball which was thrown vertically up*. At the same time, we asked expert annotators to identify the minimal referent (typically one word) in this noun phrase, which in this example would be *ball*. The basic referent must be specified in the value of the *min* attribute value. Below, we show the fully annotated instance:

- <np id="1000_1" min="ball">The red ball which was thrown vertically up</np> strikes the target first, and thus it returns the earliest.

While the above general guidelines cover many cases, there are many special cases that require special handling as exemplified below.

Special Cases :

- *Personal pronouns* : Sometimes students use in their responses first person personal pronouns such as I, me, us, my etc. For these pronouns, we assign a special id of 0 (id="0") as there wouldn't be any explicit entity/entities corresponding to such pronouns in the actual dialogue (rather, the referent is the speaker).

  Example 1 :
  Q: What can you say about <np id="1627_1" min="trajectory and speed">the trajectory and speed of the puck</np> ?
  A: I don't know.

  Example 2 :
  Q:What can you say about the motion of <np id="3582_1" min="stuntman">the stuntman</np> after he jumps ?
  A: he will take a slightly curved path

  When used, other types of personal pronouns (e.g. he, she, they etc.) refer to characters in our Physics problems such as Bill or Mary.

- *Pleonastic pronouns*: Some pronouns such as *it* can sometime be pleonastic meaning that they have no referent. Consider the sentence: *It is true that Newton's first law can be applied in such situations*. Here, the pronoun *it* doesn't refer to anything. In those case we assign an *refid*="-1" to indicate that there is no referent.

- *Communication breakdowns*: Sometimes students are not directly answering the previous tutor question. We categorize these utterances as communication breakdowns and divide them into two types: *soft* and *hard* communication breakdowns. This an important difference between dialogue and written text. Often, dialogue contains disfluencies, hesitations, abandoned utterances, interruptions or incomplete sentences.

  We call a communication breakdown *soft* when the human expert can locate the entity referred by the pronoun with some effort. An example is shown below where the tutor question is asking about what happens to the acceleration of the ball. The student's response indirectly answers the question by referring to the ball's motion. Implicitly, the ball's motion indicates something about the ball's acceleration as well, however, the student answer does not directly refer to the acceleration of the ball, which is the focus of the tutor question.

  Example 3:
  Q: What does Newton's second law tell you about the acceleration of <np id="213_1" min="ball">the ball</np>?
  A: it will continue to bounce.

  In such cases, we asked annotators to identify the best entity the pronoun can refer to, given the dialogue context. At the same time, they were instructed to preced the value of refid by * and to use the "comments" attribute of the *p* tag to indicate it as a communication breakdown.

  We call a communication breakdown *hard* if the human expert cannot infer the referent of a pronoun. To understand the concept, consider an instance below.

  Example 4:
  Q: What can you say about objects near earth upon which the only force acting is the force of gravity exerted by the earth ?
  A: it is equal

  In this example, the pronoun "it" in the student response A has no valid referent simply because the student answer makes no sense given the tutor question Q, the dialogue history, and the problem description.

- *Others*: For any other cases, assign an *refid* of "-3" in tag and describe the case using the *comments*

attribute.

4. Results

Once the guidelines were specified, five pairs of annotators were formed to annotate the instances. Initially, each pair annotated 100 instances independently. Annotators discussed and resolved the differences after the first round of annotation. We also updated the guidelines in order to reduce confusion and potentially maximize future annotation agreements. After that, each annotator annotated another set of 100 instances independently. We present below the agreement scores and an analysis of the corpus.

4.1. Agreements

To gauge agreements between annotators in each pair, we considered several parameters: the types of pronouns, the values of the referents, the locations and positions of the referents.

Location of a referent indicates where the referent is located which could be any of these locations: A, Q, H_i, or PROBLEM. The positions of a referent are the start and end word indices in its location. For instance, if a noun phrase lies between 4th and 6th words (both inclusive) in student's answer, then its location is A and the positions would be 4, and 6 respectively. Agreements are computed using kappa statistics, which measures the degree of agreement between two annotators while accounting for chance.

We map *refid* of each pronoun described in section 3.2.1. to one of the values in the set {*hasNPId*, 0, -1, -2, -3} where *hasNPId* maps to the pronouns whose *refid* is other than 0, -1, -2 and -3 i.e. to the pronouns that refer to some true referents. This allows us to check agreements based on the type of referent. Kappa statistics in this case was computed for each pair of annotators and the results are: 0.83 for the two annotators in the first pair, 0.88, 0.72, 0.81, and 0.82 for the annotators in the second, third, fourth, and fifth pair, respectively. This leads to an average kappa value to 0.81 which indicates a strong inter-rater agreement.

In the previous computations, we only considered the type of referent. We also wanted to see how many times the annotators found exactly the same referent for each pronoun. This time, we considered the entities they refer to, but we ignored their locations (among A, H_i, or PROBLEM). Thus, as long as the two annotators assigned the same value to a pronoun's referent, we considered the annotators agreed with each other. As mentioned, the values for a referent are a long noun phrase (the value of <np> tag) and a short noun phrase (the value of the *min* attribute). When we considered the short noun phrase as the value, we obtained the following agreement scores: 0.87, 0.83, 0.88, 0.83 and 0.81, respectively. For the long noun phrase, the agreement scores were 0.82, 0.65, 0.84, 0.65 and 0.74. Finally, we computed agreement considering the location and position of the referent as well. This time we obtained the following agreement scores: 0.79, 0.66, 0.80, 0.60 and 0.70 for short noun phrase; and 0.76, 0.58, 0.77, 0.58 and 0.56 for long noun phrases.

4.1.1. Disagreements Analysis

The annotation disagreements were due to several reasons. One of the reasons was *complex referents* which consists of multiple nouns or noun phrases as in the example below:

TASK: Two friends are standing over a bridge over a creek. Each friend has a stone. Sarah throws <np id="282_1" min="stone">her stone</np> straight out from the bridge so that its initial velocity is horizontal. At the same time that Sarah's stone leaves her hand, Billy drops <np id="282_1" min="stone">his stone</np> so that it falls straight down.
Q: Which stone will hit the water first?
A: they will hit at the same time.

Another source of disagreements was with communication breakdown and pleonastic pronouns. For example, for some pronouns one annotator considered it a pleonastic while the other did not.

Pronouns	Count	%
hasRef (e.g. it, he, she)	1003	78.11
personal	170	13.23
pleonastic	32	2.49
communication breakdown (Soft)	32	2.49
communication breakdown (Hard)	27	2.10
others	20	2.49

Table 3: Distribution of anaphors

Location	Count	Percentage(%)
Q	577	53.22
A	342	31.54
P	125	11.53
Q_1	28	2.6
Q_2	5	0.46

Table 4: Top five locations for antecedents

4.2. Corpus Analysis

We analyzed the 1,000 annotated instances to better understand the pronoun use by students in tutorial dialogues. As we can see from the distribution in Table 3, a great majority of pronouns (78.11%) refer to actual entities. Students also use, to a lesser degree, personal and pleonastic pronouns.

We also gathered statistics about the pronoun location - the results are shown in Table 4. It can be seen that 53.22% of students' anaphors refer to entities in the most recent tutor turn (hint/question Q). They also use pronouns to refer to entities in their response (intra-turn anaphors) which accounts for 31.54% of pronouns. The third major place for the referents is the problem description (P) which account for 11.53%. About 2.6% of the time students' pronouns refer to entities in the hint that immediately precedes the most recent hint (i.e. Q_1). Interestingly, very few pronouns has referents beyond this Q_2 hint the precedes Q_1.

Table 5 shows the frequency distribution for most commonly found pronouns. The first two pronouns (it and they) account for more than 65% of the pronouns.

5. Conclusion

In this paper, we described the DARE corpus, a resource for pronoun resolution in tutorial dialogues. Resolving

Pronoun	Count	Percentage(%)
it	658	53.47
they	153	11.94
its	120	9.37
i	61	4.76
you	55	4.29
her	36	2.81
she	34	2.65
them	21	1.63
he	19	1.48
their	18	1.40
his	17	1.33

Table 5: Most common pronouns

pronouns in student responses in tutorial dialogue plays a central role in assessing the correctness of student' responses and providing accurate feedback. To the best of our knowledge, the presented DARE corpus is the first data set specifically created to address the task of pronoun resolution in tutorial dialogues. The DARE corpus contains 1,000 unique pronoun instances taken from real student-tutor interactions. The instances were annotated manually by human experts with high agreements, making the data set ready to use in the development of advanced pronoun resolution methods for tutorial dialogues, which is part of our future work. The data set is publicly available as of this writing at: http://language.memphis.edu/nobal/AR/ .

6. Acknowledgements

This research was supported in part by Institute for Education Sciences under awards R305A100875. Any opinions, findings, and conclusions or recommendations expressed in this material are solely the authors' and do not necessarily reflect the views of the sponsoring agencies.

7. References

Bangalore, S., Di Fabbrizio, G., and Stent, A. (2008). Learning the structure of task-driven human–human dialogs. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(7):1249–1259.

Kim, J.-D., Ohta, T., Tateisi, Y., and Tsujii, J. (2003). Genia corpus?a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182.

Mitkov, R., Evans, R., and Orasan, C. (2002). A new, fully automatic version of mitkovs knowledge-poor pronoun resolution method. In *Computational Linguistics and Intelligent Text Processing*, pages 168–186. Springer.

Niraula, N. B., Rus, V., and Stefanescu, D. (2013). Dare: Deep anaphora resolution in dialogue based intelligent tutoring systems. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)*, pages 266–267.

Poesio, M. and Kabadjov, M. A. (2004). A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Proceedings of LREC*.

Poesio, M. (2004). The mate/gnome proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162.

Qiu, L., Kan, M.-Y., and Chua, T.-S. (2004). A public reference implementation of the rap anaphora resolution algorithm. In *proceedings of the Fourth International Conference on Language Resources and Evaluation*.

Rahman, A. and Ng, V. (2009). Supervised models for coreference resolution. In *Proceedings of EMNLP*, pages 968–977. ACL.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

Rus, V., D'Mello, S., Hu, X., and Graesser, A. C. (2013a). Recent advances in conversational intelligent tutoring systems. *AI Magazine*, 34(3).

Rus, V., Niraula, N., Lintean, M., Banjade, R., Stefanescu, D., and Baggett, W. (2013b). Recommendations for the generalized intelligent framework for tutoring based on the development of the deeptutor tutoring service. In *AIED 2013 Workshops Proceedings*, volume 7, page 116.

Stent, A. J. and Bangalore, S. (2010). Interaction between dialog structure and coreference resolution. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 342–347. IEEE.

Su, J., Yang, X., Hong, H., Tateisi, Y., and Tsujii, J. (2008a). Coreference resolution in biomedical texts: a machine learning approach. In Ashburner, M., Leser, U., and Rebholz-Schuhmann, D., editors, *Ontologies and Text Mining for Life Sciences : Current Status and Future Perspectives*, number 08131 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany.

Su, J., Yang, X., Hong, H., Tateisi, Y., and Tsujii, J. (2008b). Coreference resolution in biomedical texts: a machine learning approach. *Ontologies and Text Mining for Life Sciences*, 8.

Versley, Y., Ponzetto, S. P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., and Moschitti, A. (2008). Bart: A modular toolkit for coreference resolution. In *Proceedings of ACL*, pages 9–12. ACL.