Relating Student Text to Ideal Proofs: Issues of Efficiency of Expression[1]

Pamela W. Jordan [2], Maxim Makatchev and Umarani Pappuswamy
University of Pittsburgh, Learning Research and Development Center
3939 O'Hara Street, Pittsburgh PA, 15260, USA.

Abstract. In this paper we focus on text analysis issues and describe the linguistic problems we have encountered in working with an ideal two-column proof as the pedagogical foundation for the tutoring system. We discuss how the student's explanation is challenging to assess relative to an ideal two-column proof since in the explanation (1) parts of the proof can be skipped and may not necessarily reflect a gap in the student's knowledge and (2) parts of the proof can be merged. We hypothesize that skipping and merging are due to efficiency of expression and begin to examine the implications for assessing the depth of a student's understanding of the domain. Finally we outline the solutions we have attempted and are considering for these issues.

Keywords. Text analysis, Mixed-language explanations

1. Introduction

One goal of the Why2 project has been to experiment with and evaluate the effectiveness of a variety of NLP techniques for tutoring. We selected qualitative physics as the tutoring domain as it required students to generate extended natural language explanations. To facilitate this experimentation we divided the interaction into two phases, (1) entry of an essay by a student that answers and justifies the answer to a qualitative physics problem and (2) a follow-up dialogue that over an extended interaction will help the student remedy flaws in his essay. In this paper we will focus on how the Why2-Atlas system addresses the first phase.

In the most recent version of Why2-Atlas we adopted a goal of eliciting an essay that covers a subset of an ideal two-column proof, where physics principles in the right column are used as justifications of the facts in the left column. Previous versions of the systems elicited a subset of the facts in the left column [11]. A hypothesized advantage of requiring discussion of justifications is that it encourages deeper learning because physics principles are now explicitly exercised. Although so far we have collected 48 problem interactions between students and the current version of the system, we have not yet started any formal studies of students' language use. Instead, we show some excerpts

[1]This work was funded by NSF grant 0325054 and ONR grant N00014-00-1-0600.
[2]Correspondence to: Pamela Jordan, Tel.: +1 412 624 7459; Fax: +1 412 624 7904; E-mail: pjordan@pitt.edu.

Question: Suppose a man is in an elevator that is falling without anything touching it (ignore the air, too). He holds his keys motionless right in front of his face and then lets go. He neither tosses them up nor throws them down; he just releases his grip on them. What will happen to them? Explain.

Prescribed Explanation: (18f) The keys remain in front of the man's face the whole way down. We can show this by analyzing forces and motions along the vertical dimension. (12f) Before the release of the keys the man and the keys have the same velocity because they are moving together. (4f) After the release, the only forces on the man and the keys are gravitational. (6f) Thus, their net forces are equal to their gravitational forces. (5jv) Now because <gravitational force = mass * g> and (8jv) <net force = mass * acceleration>, we know that (10f) <acceleration = g> for both the man and the keys. (13) Because their accelerations are the same, and their initial velocities are the same, the man and the keys have the same final velocity in accordance with <acceleration = (final velocity - initial velocity) /elapsed time>. (14) Because their acceleration is constant and they have the same initial and final velocity, we know that the man and the keys have the same average velocity as well. (15) This implies that they have the same displacements at all time, because <average velocity = displacement / elapsed time>. (16f)(17fjq) Because the man's face and the keys start at the same height, and they have the same displacement at all times, they have the same vertical position at all times. (18) Thus the keys remain in front of the man's face during the whole trip down."

Figure 1. The statement of the problem and its prescribed essay. The numbered codes are insertions we have added that relate a sentence or clause to parts of the ideal proof.

from the corpus to illustrate the linguistic problems we have encountered in working with the two column proof as the pedagogical foundation for the tutoring system. We will discuss how the student's explanation is challenging to assess relative to the ideal proof since in the explanation (1) parts of the proof can be skipped and this may not necessarily reflect a gap in the student's knowledge and (2) parts of the proof can be merged.

First we will give an overview of Why2-Atlas. Next we describe the issues involved in evaluating a student essay relative to the ideal two-column proof in which steps may have been omitted or merged due to efficiency of expression. Finally we outline the solutions we have attempted and future work we are considering.

2. The Why2-Atlas System

Why2-Atlas covers four qualitative physics problems on introductory mechanics. When the system presents one of these problems to a student, it asks that he type an answer and explanation and informs him it will analyze his final response and discuss it with him. One of the problems Why2-Atlas covers is shown in Figure 1 along with the prescribed ideal response[1], which the student is shown before moving on to a new physics problem. An initial verbatim student response is shown in Figure 2 and the response from the same student after several follow-up dialogues with Why2-Atlas is shown in Figure 3. Each of the essays shown has numbered codes inserted for exposition purposes that we will explain at the end of this section.

[1]The domain experts simplified some aspects of the explanation and made others more complicated in order to balance exposure to principles during evaluations of the system.

(18) They will'hang in the air'in front of him. (12f-14f) Both he and the keys are falling at the same velocity, And no force is being exerted on the keys by the man, (17f) so they'll be right next to one another as they fall.

Figure 2. An initial verbatim student explanation for the problem in Figure 1. The numbered codes are insertions we have added that relate a sentence or clause to parts of the ideal proof.

(4f) The only force acting on both man and keys is gravity. (5f) The magnitude of the force of gravity on the man and the keys is its mass times g. (7f) The magnitude of the net force on each body equals its mass times g.
(8jv) <net force = mass * acceleration>. (8f) Therefore, the magnitude of both accelerations is f/m. (10f) This is equal to g.
(13jv)<Acceleration = (final velocity - initial velocity)/elapsed time>. (12f) The initial velocity of the keys is the same as the initial velocity of the man. (13f) The final velocities are also the same. The time is the same. (13jq) If the acceleration, initial velocity, and time are the same, then the final velocity is too. (15jv) <Average velocity = displacement / elapsed time> (14f) The average velocities are also the same. (17jq) If two things have the same v and t, then they have the same d.

Figure 3. A verbatim subsequent explanation from the same student in Figure 2 for the problem in Figure 1. The numbered codes are insertions we have added that relate a sentence or clause to parts of the ideal proof.

During the essay analysis phase, each sentence entered by the student is first subjected to a pre-processor that segments complex sentences, marks up equations, such as <average velocity = displacement / elapsed time>, and corrects spelling errors. Each corrected, marked-up sentence segment is then competitively analyzed by three different sentence-segment analysis techniques and a final interpretation is heuristically selected [6]. The final output of the sentence segment analysis is a function-free first-order predicate logic (FOPL) representation for each sentence segment.

Our intent in applying multiple sentence-segment analysis approaches is to use each to its best advantage relative to a particular time-slice in the life-cycle of the knowledge development effort for the tutoring system. At a given time-slice one approach may be functioning better than another for certain types of sentence segments. But since the knowledge development is on-going, we anticipate that the performance may change over time.

The selection heuristics for choosing which result to use depend on the FOPL representation language but not on the analysis techniques. The heuristics filter and rank each output representation using an estimate of whether a resulting representation either over or under represents the segment. The estimate combines counts of matches between the root forms of the words in the natural language segment and the constants in the FOPL representation returned by each method.

While any number of analysis approaches could be incorporated into the system in this way, we are currently using three that represent a range of approaches: symbolic, statistical and a hybrid. The statistical approach uses classes defined relative to the two-column proof and thus the FOPL representation for each class is stored and accessed when needed by the selection heuristics. The other two approaches directly produce an FOPL representation.

As the final step in analyzing a student's essay, an assessment of correctness and completeness is performed by matching the final FOPL representations of the student's

essay to nodes of an augmented assumption-based truth maintenance system (ATMS) [8].[2] An ATMS for each physics problem is generated off-line. Both good and buggy physics rules are applied to the givens specified in the problem statement. Each anticipated student misconception is treated as an assumption (in the ATMS sense), and all conclusions that follow from it are tagged with a label that includes this assumption along with any other assumptions needed to derive that conclusion.

Completeness in Why2-Atlas is relative to an ideal two-column proof generated by a domain expert. The ideal proof for the problem in Figure 1 is shown in Figure 4 where facts appear in the left column and justifications that are physics principles appear in the right column. Justifications are further categorized as vector equations (e.g. <Average velocity = displacement / elapsed time>, in step (15) of the proof), or qualitative rules (e.g. "so if average velocity and time are the same, so is displacement" in step (15)). A two-column proof is represented in the system as a directed graph in which nodes are facts, vector equations, or qualitative rules that have been translated to the FOPL representation language off-line. The edges of the graph represent the inference relations between the premise and conclusion of modus ponens.

Matches of input representations against the ATMS and the two-column proof do not have to be exact. Further flexibility in the matching process is provided by examining an inferential neighborhood of radius N (in terms of graph distance) from matched nodes in the ATMS to determine whether it contains any of the nodes of the two-column proof. This provides an estimate of the inferential proximity of a student's utterance to nodes of the two-column proof. Details of the analysis of correctness and completeness of the essay are provided in [8] and will not be covered further in this paper.

To determine which nodes are minimally required to be covered in a student essay, initially, our pedagogical expert for the domain estimated that for the four problems addressed by the system, a minimally acceptable essay should include all the facts (left column) and a subset of the justifications (right column). In the case of the problem shown in Figure 1, the minimum required subset of justifications is the three justifications in bold in Figure 4. The decision to require just a subset of the justifications was motivated by the purely practical need to limit the time that a student would have to spend in the worse case on a particular problem (i.e. schedule and budget constraints). Thus for this selection process we used a rule of thumb to select only those justifications that involved the most fundamental physics principles. The intuition is that not all justifications are of equal importance in the learning process.

In the remainder of the paper, we will refer to an entire proof step using its step number but to refer to parts of a step we will append to the step number "f" for the fact, and "j" for the justification. To refer to the parts of a justification we will append after the "j", "v" for the vector equation and "q" for the qualitative statement. For example, (15) refers the full step 15 in the proof, (15f) refers to just "The keys and the man have the same displacements at all times", (15jv) refers to just "<Average velocity = displacement / elapsed time>", (15jq) to just "so if average velocity and time are the same,so is displacement", and (15j) to the combination of (15jv) and (15jq).

[2] This matching step is skipped for any sentence-segment result that is a product of the statistical approach.

Step	Fact	Justification
1	The keys and the man have a gravitational force on them due to earth	If an object is near earth, it has a gravitational force on it due to the earth
2	There is no force on the keys due to air friction	The force due to air resistance is zero when there is no relative motion between air and the object
3	There is no force on the man due to air friction	The force due to air resistance is zero when there is no relative motion between the air and the object
4	The only force on the keys and the man is the force of gravity	Forces are either contact forces or the gravitational force
5	The magnitude of the force of gravity on the man and the keys is its mass times g	The force of gravity on an object has a magnitude of its mass times g, where g is the gravitational acceleration
6	The net force on each body equals the force of gravity on it	¡net force = sum of forces¿, so if an object has only one force on it, then the object's net force equals the force on it
7	The magnitude of the net force on each body equals its mass times g	Transitivity: if A=B and B=C, then A=C; if A=B and B<C, then A<C, etc.
8	The magnitude of each body's acceleration equals its net force divided by its mass	<**net force = mass * acceleration**>**, so the magnitude of the net force on an object equals its mass times the magnitude of its acceleration**
9	The magnitude of each body's acceleration equals its mass * g divided by its mass	Transitivity: if A=B and B=C, then A=C; if A=B and B<C, then A<C, etc.
10	The magnitude of each body's acceleration equals g	Canceling out: If A=B*C/C then A=B
11	The key and the man have the same acceleration	Transitivity: if A=B and B=C, then A=C; if A=B and B<C, then A<C, etc.
12	Because the man was holding the keys initially, and he moves along with the elevator, the keys and the elevator have the same initial velocity	If an object moves along with an agent, they have the same velocity, acceleration and displacement
13	At every time interval, the keys and the man have the same final velocity	<**Acceleration = (final velocity - initial velocity)/elapsed time**>**, so for two objects, if the acceleration, initial velocity and time are the same, so is final velocity**.
14	The man and the keys have the same average velocity while falling	If acceleration is constant, then <average velocity = (vf+vi)/2>, so if two objects have the same vf and vi, then their average velocity is the same.
15	The keys and the man have the same displacements at all times	<**Average velocity = displacement / elapsed time**>**, so if average velocity and time are the same, so is displacement**.
16	The keys and the man have the same initial vertical position	given
17	The keys and the man have the same vertical position at all times	<Displacement = difference in position>, so if the initial positions of two objects are the same and their displacements are the same, then so is their final position
18	The keys stay in front of the man's face at all times	

Figure 4. The ideal "proof" used in Why2-Atlas for the Elevator problem in Figure 1. The prescriptively required justifications are in bold.

3. Issues in Relating Text to Proofs

Note that while the prescriptive and subsequent student essays do include both facts and justifications, relative to the ideal proof many facts are skipped, parts of justifications are skipped and parts of multiple steps are merged. Looking at the numbers that we have inserted in the prescribed essay, note that proof parts (1f)-(3f),(7f),(9f),(11f) and (16f) are missing from the essay. Note too that some of these facts are also missing from both of the student's essays. In (13)-(15) of the prescriptive essay, the qualitative part of a justification and the fact that it helps justify are merged so that only the specific fact is mentioned. In the case of (5jv),(8jv) and (10f) in the prescriptive essay, these step components are merged into one sentence.

We will start first with skipped facts. We claim that fact (1f) is skipped in both the prescriptive and student essays, because saying the fact (4f) pragmatically presupposes the fact (1f). Although there is no settled formal definition of a pragmatic presupposition, loosely, it is any background assumption that arises for a statement [7]. For example, a presupposition of (4f) is "there exists a force due to gravity on the keys" because the definite noun phrase is thought to act as a presuppositional trigger [7].

Furthermore, we claim that explicitly stating both (4f) and (1f) in natural language is in violation of Grice's Maxim of Quantity (say no more than is necessary) [4] and thus it should be unlikely for both to appear together in a written explanation. Grice hypothesized four Maxims of Conversation: (1) Quality; say only what you believe to be true and have evidence for, (2) Quantity; say just the right amount relative to the purpose of the exchange, (3) Relevance; be relevant, (4) Manner; be brief, orderly and avoid obscurity and ambiguity.[3] He also hypothesized that seeming violations of these maxims would lead a hearer to seek an interpretation that would resolve the violation, thus leading to conversational implicatures. We claim these maxims and responses to violations are relevant for text as well given that text can be viewed as a variant of conversation [2].

We see that students do clearly violate some of these maxims. For example, during dialogues with Why2-Atlas, we saw many cases of students violating the Maxim of Quality. When the system requested that they provide evidence for an answer they just gave, students sometimes admitted they guessed at the answer. But what, if any, useful conversational implicatures the tutor should make for seeming violations remain to be determined. In the case of a violation of the Maxim of Quantity, in which more than is necessary is said, it may be a helpful diagnostic of the student's grasp of the material. Including what is typically viewed as unnecessary redundancy could be interpreted as the student failing to see the obvious connections or failing to know the premises for a rule. So it may be best to treat what should be pragmatic presuppositions as optional and their inclusion in a text is perhaps indicative of a student who needs help in achieving a more coherent picture of the reasoning process.

Facts (2f) and (3f) are also typically left out of student essays because both are givens and explicitly indicate that a force is to be ignored. Because this is near the beginning of the problem, saying (4f) means it is likely the given information was still salient for the student and that if the student considered all types of forces then (2f) and (3f) were used in reasoning. Another given (16f), appears in the prescriptive essay as part of (17q). Here the problem statement is no longer as salient as at the beginning of the text. Finally, we

[3] Although the maxims provide insight into why particular content is made explicit or not, they offer little guidance for implementation.

note that (7f), (9f) and (11f) each are conclusions of simple math manipulations and thus their inclusion could be a signal that the student believes these are of equal or greater importance than the other steps. Work in natural language generation operationalizes related omissions of intermediate reasoning and parts of rules used during reasoning by appealing in part to a model of the user's knowledge and attention [5].

Looking next at merged parts of the proof, we claim that this is related to Grice's Maxim of Manner (brevity) and again appeal to work in natural language generation to explain why this efficiency of expression is a possible display of brevity. A goal-directed view of sentence generation suggests that speakers can attempt to satisfy multiple goals with each utterance [1] and thus the same form can opportunistically contribute to the satisfaction of multiple goals [10]. But there are trade-offs across linguistic levels so that an intention which is achieved by complicating a form at one level may allow the speaker to simplify another level by omitting important information. (E.g. a choice of clausal connectives at the pragmatic level can simplify the syntactic level [3]).

Within a language generation system overloading is accomplished in part during the *sentence aggregation* process [9]. Sentence aggregation optionally combines simple sentences into more complex ones and often improves coherency of a text. So we claim that aggregation may reflect coherency in the mind of the student and that short, choppy, non-aggregated sentences may indicate the student is having trouble grasping the material. However, the student may avoid aggregation if text analysis frequently fails to understand his aggregated sentences. Note that the initial student essay is more aggregated than the subsequent one and that the short sentences are very similar to the corresponding part of the two-column proof. The system often bottoms-out and explicitly tells the student what is was trying to get him to add to his essay.

We also see semantic aggregation as well, in that specific and generic content are frequently merged so that we see the more specific content more than just generic (as in (17jq) in Figure 3) or a combination of generic and specific (as with (5jv),(8jv) and (10f) in Figure 1).

4. Initial Solutions and Future Work

With respect to which parts of the proof can be skipped in students' explanations, all parts of the ideal two-column proof are represented within the system but just those parts that are minimally required are marked as required. Thus the student is permitted to skip any part of the proof that is not marked as required. The dialogue then only addresses the marked nodes of the proof that could not be linked to the student text. While deciding what to mark as required, we made judgments about when steps are presupposed and only marked facts that are not presuppositions or salient givens. We did still mark simple math conclusions such as (10f) as required but in retrospect perhaps should not have.

With respect to parts of the proof being merged in the student essay, the system must handle one-to-many mappings between a sentence and parts of a proof. We currently have no mechanism to distinguish acceptable and unacceptable aggregations and perform only an assessment of completeness by matching representations of what the student said to an ideal proof and the ATMS. A problem with this matching process arises if the selected sentence-segment analysis is a product of the statistical approach. Recall that the statistical approach classifies a sentence-segment according to classes that are

the nodes of the proof graph representation. But the statistical approach currently returns just the single most likely class. Thus when sentence segmentation fails and steps are merged in the sentence, part of the content of the sentence will be missed and will not be matched to the proof. This is not a problem for the other two sentence-segmentation approaches. However, currently the output of the statistical approach is the result that is most frequently selected by the heuristics. Thus it is worthwhile investigating whether we can automatically recognize contexts in which this approach should return multiple classifications in order to see if it will improve the accuracy of linking sentences to the two-column proof.

Further, while the sentence-segmenter may break sentences into clauses appropriate to parts of the proof representation, it was not specifically developed to support segmentation particular to the two-column proofs. It was instead developed to improve parsing efficiency by breaking complex sentences into simpler ones. We are considering retraining the sentence-segmenter relative to the parts of the proof as this should still accomplish its original purpose and potentially improve the ability to handle one-to-many mappings.

Finally, we will consider the extent to which proof transformation methods used in NL generation [5] can be effectively reversed for analysis purposes.

References

[1] Douglas E. Appelt. Some pragmatic issues in the planning of definite and indefinite noun phrases. In *Proceedings of 23rd ACL*, 1985.

[2] Herbert H. Clark. *Using Language*. Cambridge University Press, Cambridge, England, 1996.

[3] Barbara Di Eugenio and Bonnie Webber. Pragmatic overloading in natural language instructions. *International Journal of Expert Systems, Special Issue on Knowledge Representation and Reasoning for Natural Language Processing*, 9(1):53–84, March 1996.

[4] H. Paul Grice. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 64–75. Academic Press, New York, 1975.

[5] Helmut Horacek. Generating inference-rich discourse through revisions of RST-trees. In *Proc. of AAAI*, pages 814–820, 1998.

[6] Pamela W. Jordan, Maxim Makatchev, and Kurt VanLehn. Combining competing language understanding approaches in an intelligent tutoring system. In *Proceedings of the Intelligent Tutoring Systems Conference*, 2004.

[7] Stephen C. Levinson. *Pragmatics*. Cambridge University Press, Cambridge, England, 1983.

[8] Maxim Makatchev and Kurt VanLehn. Analyzing completeness and correctness of utterances using an ATMS. In *Proceedings of Int. Conference on Artificial Intelligence in Education, AIED2005*. IOS Press, July 2005.

[9] Ehud Reiter and Robert Dale. Building applied natural-language generation systems. *Journal of Natural-Language Engineering*, 3:57–87, 1997.

[10] Matthew Stone and Bonnie Webber. Textual economy through close coupling of syntax and semantics. In *Proceedings of 1998 International Workshop on Natural Language Generation*, Niagra-on-the-Lake, Canada, 1998.

[11] Kurt VanLehn, Pamela Jordan, Carolyn Rosé, Dumisizwe Bhembe, Michael Böttner, Andy Gaydos, Maxim Makatchev, Umarani Pappuswamy, Michael Ringenberg, Antonio Roque, Stephanie Siler, and Ramesh Srivastava. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proceedings of Intelligent Tutoring Systems Conference*, volume 2363 of *LNCS*, pages 158–167. Springer, 2002.