Integrating Geometrical Pattern Recognition With Natural Language Question Answering

John F. Sowa, Arun Majumdar, and William Full

In December 2004, the DARPA Information Processing Technology Office solicited suggestions for "Grand Challenge" problems that could be used to evaluate progress in artificial intelligence. They were looking for problems that could be appreciated by a lay audience with no knowledge of AI technology, could be solved with existing technology for simple examples, but had sufficiently difficult extensions or ramifications that they would challenge the capabilities of AI technology for the next 10 to 20 years. The task we suggested is one that nearly every two-year-old child solves: the problem of learning to integrate visual, tactile, and motor information with language. To evaluate progress on this task, we proposed that any AI research group that wished to respond to the challenge be given a collection of binocular pictures, still or moving, together with some natural-language questions about those pictures. Any AI system they develop would be asked to determine which pictures could answer any of the questions and to state those answers.

This problem, which any normal child can solve, involves critical aspects of both human and animal intelligence. The higher mammals, especially the primates, learn to integrate visual, tactile, and motor information in infancy, yet they never get to the level of using a language with symbolic references to the things and actions they see and do. Human infants combine a similar ability to integrate geometrical information from multiple sources with a remarkable ability to relate it to language. Both aspects of this problem, the geometrical and the linguistic, have been solved in simple cases with existing AI technology, but none of the existing geometrical technology is as good as a typical mammal's, and none of the linguistic technology is as good as a typical three-year-old human's. The following sentence, for example, was spoken by a child at age 2 years and 10 months: *When I was a little girl, I could go "geek-geek," like that, but now I can go "This is a chair."*

As stated, this problem does not presuppose any theory of linguistics, any knowledge representation language, any method of computational geometry, or any theory of how the brain works. Many researchers in AI and other branches of cognitive science have been developing such theories, but most of them have been working on isolated aspects of the problem without considering how they are integrated in a child or in a robot that might be able to simulate the abilities of a child. The purpose of this challenge is to focus attention on the problem of integrating language with sensory-motor processing without favoring any particular theoretical approach.

DARPA IPTO Grand Challenge Proposal

Submitted by: John F. Sowa, Arun Majumdar, and William Full

One liner: Integrating geometrical pattern recognition with NL question answering.

Brief explanation: People commonly use a seamless integration of linguistic and contextual references in talking about physical situations, but most research in natural language processing has had little contact with research in robotics and scene recognition. The purpose of this proposal is to encourage collaboration and integration of the research and development in both areas and to challenge the collaborators to demonstrate the benefits for NLP, robotics, and scene recognition. The practical results would have dual-use benefits in applications ranging from military surveillance and battlefield analysis to travel, child care, and business security. The same technology for answering questions about geometrical patterns in dynamically changing physical situations could also be applied to questions about intrusions and bottlenecks in computer networks.

Integrating geometrical analysis with NLP helps to focus attention on relevant details. For a question about a missing child wearing a red shirt, the details about buildings, adults, and scenery need not be analyzed in detail unless something resembling the child happens to be present. For interpreting language, information from the scene can help resolve many ambiguities about pronouns, noun modifiers, verb arguments, and the attachment of modifying phrases and clauses. For analyzing computer networks, an insightful question can focus attention on relevant features and subnets in a world-wide system. Geometrical analysis benefits NLP by providing context, and NL questions benefit geometrical analysis by focusing attention on relevant features.

The basic challenge would be to address the end-to-end problem of linking the two extremes of scene recognition and natural language understanding. Since very few groups have expertise in the entire range of technologies that are needed, the overall task could be divided into four steps:

1. Analyze an image and generate a geometrical model of the scene.

2. Map the geometrical model to a description in some knowledge representation language.

3. Map the English question a logical form in some knowledge representation language, which could be the same language as in step #2 or some superset of it.

4. Answer the question from step #3 in terms of the description from step #2.

These steps may be performed iteratively to allow feedback from one step to assist in the interpretation of the other steps. Such feedback normally occurs in human vision and language understanding, and it could be valuable for focusing attention in image analysis and for providing context for language understanding. Since image recognition groups

are usually distinct from the NL groups, they could work separately or in partnership. It's also possible that multiple NL groups might work with the same image recognition group, or multiple image recognition groups might work with the same NL group.

Some care would be needed to choose appropriate examples for the development stages and for the evaluation stages. Ideally, all the ontology of geometrical patterns and all the vocabulary of NL terms would be represented in the sample data given to the developers. In the evaluation stage, similar images and vocabulary would be used, but a learning stage could be added that would allow systems to learn new patterns and the words for them.

Static images would probably be easier to analyze and describe than dynamically changing movies, but the latter might be more realistic and more important for many of the most significant applications. Some examples of both static and dynamically changing images might be included in both the examples for the development stage and the test cases for the evaluation stage.

Research technology checklist

Which of the following technologies are addressed by the Grand Challenge proposal?

x_learning
x_knowledge representation
x_reasoning
x_perception
x_multi-modal interaction/human-computer interaction
x_natural language processing
__other (please list)

Other remarks: Even without learning, the end-to-end test would be challenging, but it could easily be extended to include a learning phase. If new vocabulary and image patterns were included in the evaluation stage, the computer system might be allowed to carry on a dialog with the users to learn new words and identify new geometrical patterns.

Since there are four separate steps to the task, it is possible for different groups with different capabilities to work on any of the steps independently. Those groups that were successful in working by themselves might later form a partnership with other groups that had expertise in one or more of the other steps. In the evaluation phase, groups could be scored on each of the four steps. There could be separate winners in each step as well as an overall winner for the best end-to-end solution (and that winner might be a partnership of two or more groups working on different parts of the task). The most challenging and potentially the most important approaches would involve tightly integrated analyses that used information from the questions to focus attention on significant features of the scenes and used information from the scenes to resolve ambiguities in the language.

Criteria checklist

How does the proposal rate against IPTO criteria? Use '+,' '-,' or '?'

1. Clear & compelling demonstration of cognition	
+	a. The test should be a proxy for problems requiring cognitive capabilities.
+	b. The test should not be "game-able" or solvable by "cheap tricks"
+	c. It should not be solvable by brute force computation, alone, and it should not lend itself to idiot savant solutions
+	d. The test should require integration of multiple cognitive capabilities. It is desirable that the portfolio of tests includes sensing and acting (i.e., situated cognition).
2. Clear & simple measurement	
+	a. The test should have a clear & simple measure for measuring success.
+	b. The test should specify what must be done, not how to do it.
+	c. It is desirable to have a graduated sequence of increasingly more difficult problems.
?	d. It is desirable to have tests that are automatically score-able.
+	e. It is desirable that the tests be easy to create and run and that test results be reproducible.
3. Decomposable & diagnostic	
+	a. The test should be decomposable into sub-tests or sub-measurements for different aspects of cognition.
+	b. The test should be diagnostic (failure to pass the test should point the way to future improvements).
+	c. It would be desirable to have partial, intermediate results (scores are not just "Pass/Fail").
4. Ambitious & visionary, not unrealistic	
+	a. It should not be a toy problem
+	b. It should represent technical/scientific goals achievable within a 10-20 year window.
+	c. It should not be something that a computer can already do.
+	d. It is desirable to have military relevance (eventual).
5. Compelling to public	
+	a. It should be simple to explain and convey to the general public.
6. Motivating for researchers	
+	a. It should generate enthusiasm in the research community.
+	b. It is desirable to have a low cost of entry so that work on the problem can begin right away.
+	c. It is desirable to enable continuous testing, perhaps over the web.

The question mark for part 2.d indicates that some versions of the tests could be scored automatically while more complex versions might require human evaluation. For example, the question "Which of the following pictures show a child walking with an adult?" could be scored automatically, but an open-ended question such as "What is the boy in this picture doing?" might be harder, but not impossible to score automatically.