Abductive Metareasoning for Truth-Seeking Agents

Joshua Eckroth
Department of Computer Science
The Ohio State University
Columbus, Ohio, USA

My research seeks to answer the question of how any agent that is tasked with making sense of its world, by finding explanations for evidence (e.g., sensor reports) using domain-general strategies, may accurately and efficiently handle incomplete evidence, noisy evidence, and an incomplete knowledge base. I propose the following answer to the question. The agent should employ an optimal abductive reasoning algorithm (developed piece-wise and shown to be best in a class of similar algorithms) that allows it to reason from evidence to causes. For the sake of efficiency and operational concerns, the agent should establish beliefs periodically rather than waiting until it has obtained all evidence it will ever be able to obtain. If the agent commits to beliefs on the basis of incomplete or noisy evidence or an incomplete knowledge base, these beliefs may be incorrect. Future evidence obtained by the agent may result in failed predictions or anomalies. The agent is then tasked with determining whether it should retain its beliefs and therefore discount the newly-obtained evidence, revise its prior beliefs, or expand its knowledge base (what can be described as anomaly-driven or explanation-based learning).

When the agent is considering whether its failed predictions or anomalies are the result of false beliefs or limitations in its knowledge, or instead the result of incomplete or noisy sensor reports, the agent is performing a kind of metareasoning, or reasoning about its own reasoning (Schmill et al. 2011). My approach treats this metareasoning procedure as itself abductive. When faced with failed predictions or anomalies, the agent attempts to explain its potential failure of reasoning. Possible explanations are that the agent committed to incorrect beliefs based on prior misleading evidence. Or, the newly-obtained evidence is misleading and the agent does not possess incorrect beliefs. A further explanation is that the agent's knowledge base is incomplete, and that the anomaly resulted from the agent not having the proper facts about what kinds of events are possible in the world. The abductive metareasoning procedure (which utilizes the same abductive inference algorithm as the first-level reasoning procedure) produces its best explanation. Based on this explanation, the agent may attempt to repair its beliefs, ignore the newly-obtained evidence, or expand its knowledge base. These "fixes," such as expanding its knowledge base, may themselves be reverted if the agent reaches further failed predictions or anomalies in the near future.

Methodology

I claim that certain design choices in the first-level abductive reasoning strategy and the abductive metareasoning strategy make for more efficient and accurate truth-seeking agents. A more efficient agent naturally requires fewer resources (such as time) to do its job. A more accurate agent commits to more true beliefs about the world and expands its knowledge base with more relevant and truthful content.

Demonstrating that certain design choices result in improved performance requires that two agents are simulated in the same task domain with the same world events. Only when the two agents are measured against the same "truth" can their performances be compared. After repeated experiments in which the world or "truth" varies, a statistically-significant difference between one set of design choices and another may be established.

Furthermore, in order to support the claim that certain reasoning and metareasoning strategies are domain-general, and consistently performant across a variety of domains, the agents must be able to perform different kinds of tasks from different domains.

I have developed a software testbed that facilitates experiments with these requirements. The testbed has several important features. First, the testbed separates the reasoning and metareasoning systems from the domain-specific systems. This allows changing the domain-specific task without changing the reasoning strategies. Second, two agents with different properties can be simulated over a wide variety of worlds that vary based on parameters and controlled random variations. Third, the testbed automatically produces statistical and graphical analyses that allow researchers such as those in my research group to quickly ascertain whether an experiment was effective. Furthermore, the testbed enables drill-down from large experiments across many parameter variations to single cases. A graphical interface supports close scrutiny of an agent's reasoning processes on a particular task.

Task domains

The software testbed provides two different tasks in which to test various reasoning strategies. The first task is a simula-

tion of video surveillance. Entities (which all have the same shape but may differ in color) move somewhat randomly throughout a space that is observed by simulated video sensors. Some sensors are able to report the colors of entities, others are not (entities observed by these sensors appear gray). The agent's goal is to keep track of the various entities and maintain their identities. This task becomes more difficult as the number of entities in the scene increases and fewer sensors are able to detect color. Lack of color in sensor reports causes the agent to have access to incomplete information about the world.

The second task is Chinese word segmentation. The agent reads a sequence of Chinese characters; these characters are not segmented into words (the spaces have been removed from the test data). The goal is to segment the characters into words, based on knowledge about Chinese word frequencies, word transitions, and so on. Noise is simulated in this domain by swapping a random subset of the characters to other random characters. Some of the words are "out-of-vocabulary" words, meaning they were not present in the training set that the agent used to build its knowledge base. Thus, some words may need to be learned. Incomplete information is also possible because the agent does not always receive a complete sequence of characters, but only a subset, which may contain portions of true words. The task becomes more difficult as noise increases, the agent's knowledge base is more limited (which produces more out-of-vocabulary words in the test data), and smaller subsets of the character sequence are available.

The video surveillance domain tests agents on their ability to handle incomplete information. The Chinese word segmentation domain tests agents on their ability to handle incomplete information, limited knowledge, and noise, and any combination of these factors.

Preliminary results (as of April 2012)

The first-level abductive reasoning process, with no metareasoning facility, has proven to be effective. In both task domains, agents achieve relatively high precision and recall. F-scores for the Chinese word segmentation task are presently around 0.94, while the state-of-the-art scores, on the same datasets, are about 0.97 (Zhao, Huang, and Li 2006; Emerson 2005). The video surveillance domain is synthetic and thus difficult to compare to other systems. However, I have reason to believe the first-level abductive reasoning process is sufficiently capable and implemented properly.

The abductive metareasoning process is still undergoing considerable development. I have tested its ability to correct mistaken prior beliefs by reconsidering past evidence in light of new evidence, and its ability to invoke learning. Experiments indicate that the belief correction strategy increases accuracy but also increases resource usage because prior evidence is being reconsidered. Learning as a "fix" for anomalies appears to be working as well, although not to a degree that matches the state-of-the-art. In the Chinese word segmentation domain, my software is achieving out-of-vocabulary recall rates of about 0.35 (for the same F-score cited above) while the recent state-of-the-art achieves 0.77 out-of-vocabulary recall (Kruengkrai et al. 2009).

Proposed timeline (to July 2012)

My candidacy examination will be scheduled for early Summer. After this time, I will document and seek publication of results relating to the Chinese word segmentation task.

Individual contributions

All of the software implementation and experiments have been my own. The abductive inference strategy that these agents utilize is a modification of a strategy developed by my advisor (Josephson 2000).

References

Emerson, T. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, volume 133. Jeju Island, Korea.

Josephson, J. R. 2000. On the proof dynamics of inference to the best explanation. *Cardozo Law Review* 22:1621–1643. Reprinted in MacCrimmon, M. T. and Tillers, P., eds. 2002. *The Dynamics of Judicial Proof: Computation, Logic, and Common Sense*. Physica-Verlag. 287–306.

Kruengkrai, C.; Uchimoto, K.; Kazama, J.; Wang, Y.; Torisawa, K.; and Isahara, H. 2009. An error-driven word-character hybrid model for joint chinese word segmentation and pos tagging. 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*, 513–521. Association for Computational Linguistics.

Schmill, M. D.; Anderson, M. L.; Fults, S.; Josyula, D.; Oates, T.; Perlis, D.; Shahri, H.; Wilson, S.; and Wright, D. 2011. The metacognitive loop and reasoning about anomalies. In Cox, M. T., and Raja, A., eds., *Metareasoning: Thinking about Thinking*. The MIT Press. chapter 12, 183–200.

Zhao, H.; Huang, C. N.; and Li, M. 2006. An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, volume 1082117. Sydney: July.