Kalpa Publications in Computing

Volume 4, 2018, Pages 159–168

28th International Workshop on
Principles of Diagnosis (DX'17)

Diagnosing Alternative Facts

Johan de Kleer, Matthew Klenk, and Alexander Feldman

Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304 USA
corresponding author: dekleer@parc.com

Abstract

This paper presents an approach to applying model-based diagnosis to the task of interpreting information from a wide variety of sources: text, video, meta-data, audio, etc. Much of the information contained in the sources is contradictory, incomplete, purposely deceptive or biased. People make critical decisions based on such murky information. By automating the construction of alternatives, we can design systems that support intelligence analysts and ordinary citizens in understanding the world. We have developed a preliminary version of our HCDX tool (hypothesis construction through diagnosis). We plan to distribute this tool as open source.

1 Introduction

We live in a world where each of us has access to unprecedented amounts of information. Unfortunately, much of that information is contradictory, incomplete, purposely deceptive or biased. From the literature on confirmation bias we know that people typically commit to one interpretation of the information and ignore contradictory evidence. This approach leads everyone astray. Populaces become more polarized, communication between people with different interpretations breaks down, and analyst's ability to interpret world events becomes potentially flawed. The fundamental challenge is the tendency to commit to one ground truth with all other alternatives being invisible. Worse, even when people see multiple alternatives, they generate them as biased interpretations of other's intents. This paper proposes an initial application of model-based diagnosis [1] to the problem of hypothesis construction from data. We call our approach HCDX - hypothesis construction through diagnosis. One key advantage of HCDX, over existing approaches is that it leverages ideas from Model-Based Diagnosis to systematically and in a logically unbiased way generate all substantially different hypotheses from a knowledge base.

Although at first sight interpretation construction across multiple documents seems very distant from model-based diagnosis, under closer examination they are surprisingly similar:

- Both tasks are partially observable — all information is indirect.
- Ground truth may never be known.
- A primary goal is to determine the possible interpretations/diagnoses consistent with the evidence.

- Both can have large scale.
- The results of analysis are very important to users. Faulty systems have high cost. Faulty analyses lead to very bad decisions.
- In both domains, gathering of additional evidence to differentiate among interpretations/diagnoses is often very important.
- Both approaches require carefully designed metrics to evaluate most preferred interpretations/diagnoses.
- Both are probabilistic.

There are some important differences. In diagnosis, we seek to replace malfunctioning components. Replacing source documents or logical statements makes no sense. In diagnosis, we assume the world is random. In interpretation construction, opponents are outrightly trying to deceive us and all observers have systematic biases. For HCDX, the majority of what corresponds to the model in diagnosis, is in another module which we need to interact with through APIs.

2 Background

This research is in the context of a forthcoming DARPA program[1]. Figure 1 illustrates the phases of analysis. Analysis starts with raw material which may be video, images, meta data, speech or text. These are analyzed by various techniques to build a model of the content of each individually. Then all the individual content is merged into a common semantic representation which represents the knowledge in the source material. This knowledge may be incomplete, errorfull, deceptive or contradictory. We apply MBD to analyze this knowledge to develop diverse hypotheses for what is actually taking place in the scenario.

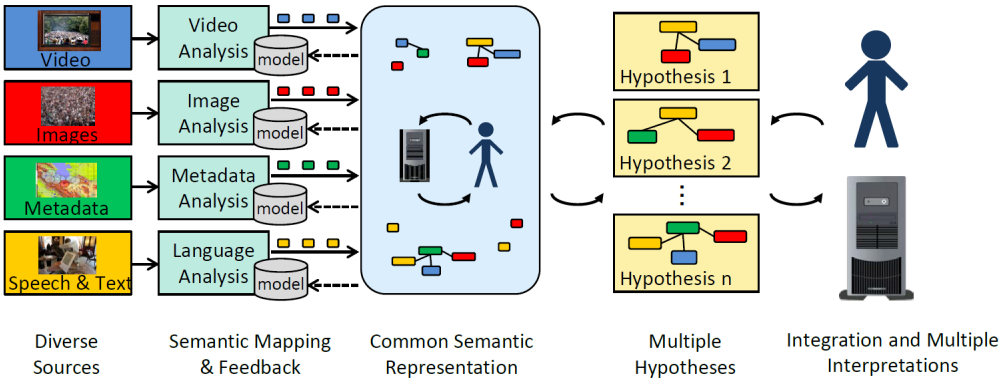

Figure 1: Analysis Phases

The knowledge is represented as Resource Description Framework (RDF) triples[2]. Consider, for example, the sources of knowledge represented in Figure 2. The question posed is "Is Russia engaging in military action in eastern Ukraine?" The raw information is contradictory and confusing. Is there

[1] http://www.darpa.mil/program/active-interpretation-of-disparate-alternatives
[2] https://en.wikipedia.org/wiki/Resource_Description_Framework

Figure 2: Scenario Sources

one, two, or three tanks. Is the observer failing to distinguish between a T72 tank vs. a T64 tank? The Russian army does not use T64 tanks. Figure 3 illustrates a simplified RDF triple representation of a possible scenario.

3 Encoding Common Semantic Representation in an ATMS based MBD

The proposed approach is to build a mapping from RDF triples to an Assumption-Based Truth Maintenance System [2] (ATMS) data base upon which we can do diagnosis. By framing hypotheses generation as a diagnostic problem, any method for assembling the scenario model must determine (1) What are the assumptions?, (2) What are the implications?, and (3) What are the exclusions? The basic element is a node, which represents a proposition or an entity (e.g., an event) from the common semantic representation. Using the common semantic representation, these nodes are connected through implications and mutual exclusions. These are all ATMS primitives that we use to implement HCDX:

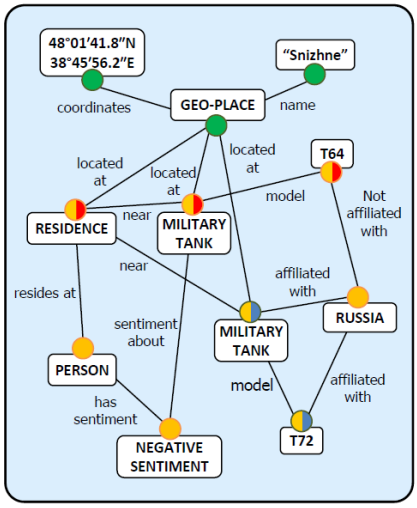

Figure 3: Common semantic representation of a scenario

- Assumptions: n. Nodes that may be included in hypotheses.
- Implications: $n_1 \wedge n_2 \wedge ... \rightarrow n$. The consequent logically follows from the antecedents (e.g.,"If the tanks is a T64, then it is not part of the Russian army.")
- Exclusions: $n_1 \wedge n_2 \wedge ... \rightarrow \perp$. The combination of propositions cannot occur, (e.g., "Rebels bombed the buildings as opposed to FSB bombed the buildings.")
- Disjunctions: $n_1 \vee n_2 \vee ...$ At least one of n_i must be true.
- Probability: $p(n_1)$. Every one of the above may have a prior probability (or confidence) associated with it.

The ATMS performs focused abductive reasoning maintaining for each node a set of minimal sets of assumptions that support the logical derivation of the node. Every combination of assumptions which leads to a contradiction (e.g., the tank cannot be both a T64 or a T72) is represented as an exclusion. In MBD, a diagnosis is expressed as the set of assumptions (correctly working components) which are consistent. A diagnosis which has more working components is usually preferred (i.e., when there is no direct evidence a particular component is faulted assume it is not). Diagnoses correspond to HCDX hypotheses in the following way. The set of all nodes supported by a diagnosis corresponds to an HCDX hypothesis. Thus the same set of nodes and edges in the ATMS representation can represent exponentially many hypotheses. There is no need to produce copies of our structure for each hypothesis. We provide a mapping from diagnoses to the corresponding knowledge elements in the knowledge base such that users can request the complete hypothesis represented as RDF triples. We refer to these diagnoses as hypothesis kernels as they only represent the assumptions of an hypothesis from which all other included nodes can be inferred. This encoding will be functional and invertible. Every knowledge base can be encoded into a unique ATMS representation. Every ATMS representation can be converted back to the identical knowledge base.

Figure 4 illustrates the mapping from knowledge elements represented as RDF triples into our ATMS-based representation. Each RDF node is mapped into an ATMS node represented by a circle

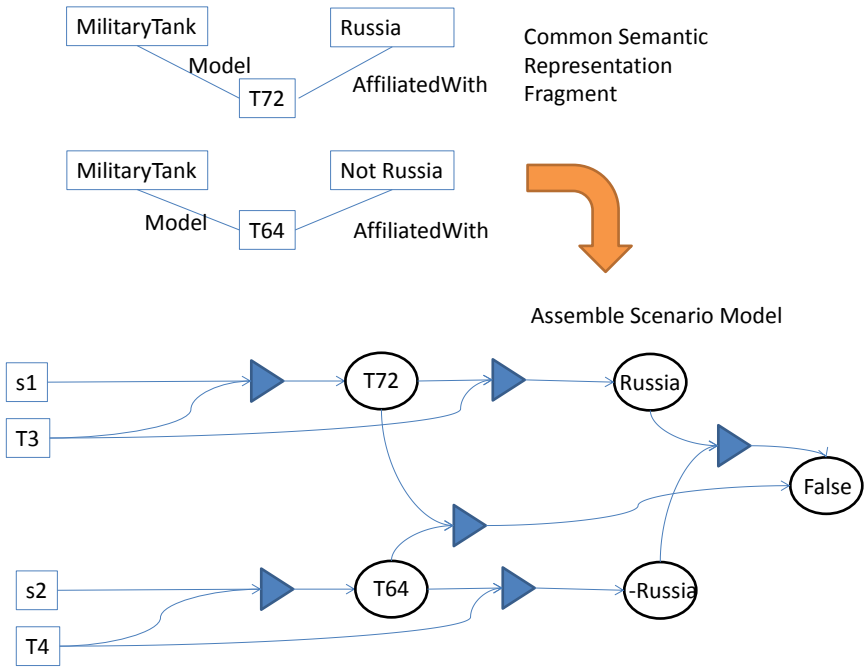

Figure 4: Document-based assembly creates assumptions for each document, implications between triple elements, and exclusions from RDF predicates.

or rectangle. Each rectangle is an ATMS node which represents an assumption. Each solid triangle represents an implication — antecedent nodes connect to flat side and a vertex connects to the consequent. s(MilitaryTank, n) is an assumption that represents an unbound entity (skolem in logic, an entity which can be linked to any other of the same type), all the other nodes represent constants. The assumptions for a skolem class such as MilitaryTank represent whether the respective entities are linked. We have also added implications to represent the usual meaning of predicates "Model" and "AffiliatedWith". Each implication that comes from a document depends on an assumption representing that document. Each skolem representing MilitaryTank is assigned a symbol "sn". (This mapping does not cover all the cases that can arise in knowledge elements, and we expect to develop far more sophisticated mappings in course of the project.)

Once the knowledge elements are mapped to the ATMS, the algorithms for interpretation construction used in model-based diagnosis can construct diagnoses which will then be refined to hypotheses. The ATMS discovers one contradiction (or nogood): T3,T4,s1,s2. This represents the conclusion that the two military tanks in the two documents cannot refer to the same entity when both T3 and T4 are believed. There are 4 maximally consistent sets of assumptions (or diagnoses):

[T3,s1,s2]
[T4,s1,s2]
[T3,T4,s1]
[T3,T4,s2]

These correspond to the 4 diagnoses:

D1: T3 only is believed. (The tank is a T72)

D2: T4 only is believed. (The tank is a T64)

D3: T3 and T4 are believed but s1 and s2 are different. (There are two different tanks).

D4: Duplicate of D3. (A consequence of our simple encoding.)

Hence there are 3 distinct maximal hypotheses. The entire example of Figure 3 can be mapped with this approach. More sophisticated mappings are required for other predicates and directly connected skolems. Notice that there are many other possible hypotheses, e.g., the empty hypothesis [] is always consistent.

4 Extending the basic encoding

While powerful, the encoding from the previous section does not incorporate all of the inferences available. In this section, we discuss three extensions to constrain the ATMS model.

4.1 Using RDF Implications

As the common semantic representation language is developed, we will analyze what kinds of relationships are extracted from the data streams. This analysis will support our second model assembly approach based on implications. In this approach, each RDF entity is a knowledge element (KE) that may be included in the explanation. Assumables are KEs for which there are no implications, and logical connections are defined in terms of the common semantic representation language.

To determine which KEs are relevant, we will initialize the scenario model with the focus node(s) of the scenario (e.g., the tank), as a KE(s) that must be included in the hypotheses. We will then identify all causally connected and mutually exclusive KEs creating additional nodes, implications, and mutual exclusions for them. Causal connections are determined by predicates in the ontology (e.g., perpetratedBy, motivates, resultsIn, justificationFor) and any background knowledge axioms used by TA2. This process terminates when all of the KEs that are causally connected to the focus have been represented. Finally, we will consider all nodes that are not consequents in any implication are assumables.

4.2 Induction Model Extension

Even if all relevant KEs are in the data stream, there is no guarantee that all relevant KEs will be correctly identified by TA1 and represented by TA2. This corresponds to the case in MBD where the model is not completely correct. Given the examples of entities, relationships between, and user's preferred hypotheses from multiple evaluations, we will learn models to infer additional relations and KEs. To identify new implications and exclusions, HCDX will draw on our work on link prediction over large triple stores [3]. To extend the model with new entities as well as new relationships, we will use analogical graph-matching techniques to map previously created hypotheses onto new models (or other parts of the same scenario). In previous work, we used this technique to infer entities that support model-based reasoning [4]. For example, if the KB did not include any information about ethnic Russians in Chechnya, HCDX could infer it if the system had previously generated hypotheses about Russian involvement Georgia and Ukraine where a similar dynamic was at work. Hypotheses that rely on induced nodes and links will suffer a penalty to their confidence metrics

4.3 Probabilistic Inference

The prior probability of a hypothesis can be computed from the priors of it assumables. In the base case there are two assumable types: those originating from source documents and those originating from entity linking. In the simple case where the source documents are independent, this component of the

probability is simply the product of the prior probabilities of the correctness of the documents in the hypothesis multiplied by one minus the priors of the documents not included in the hypothesis. The second contribution to the prior of an hypothesis is more complex to compute: what is the probability that multiple skolems refer to the same entity? We will develop specialized routines that estimate the probability that two entities refer to the same real world object. For example, the probability that two entities refer to the same tank is extremely related to the differences in the geolocations within a particular time window.

In later phases we will allow individual RDF triples to be assumables and therefore have probabilities. However, this will scale the complexity of hypothesis probability evaluation.

The ATMS is capable of exact Bayesian inference, but we expect that approach will be intractable and unnecessary for real-world problems. There simply will be far too many possible hypotheses to consider. Instead, we compute the relative probabilities of the most likely hypothesis. To determine the true probability of an hypothesis requires exact Bayesian inference. We will adapt algorithms from MBD to compute the relative order of the most likely hypothesis.

5 Evaluating HCDX

To develop HCDX, it is necessary to define metrics over which we can assess system performance and explore possible user experiences.

5.1 Semantic Coherence Metric

The semantic consistency of the hypotheses produced by HCDX should at least have the following six properties (all assume the common semantic knowledge base is complete and correct):

- The hypothesis are expressible in RDF triples extant in the original KB. Spurious triples must not be added, and if one node of a triple is present, all 3 should be.
- The hypothesis does not contain a logical inconsistency, e.g., a tank cannot have two model numbers. Avoid sins of commission.
- HCDX could generate all hypotheses given enough resources, i.e., no hypothesis is spuriously ruled out by an incorrect contradiction. Avoid sins of omission.
- Hypotheses commit to all the knowledge. All KEs in TA2 that are related to the hypothesis under consideration are either included or excluded from the hypothesis.
- Knowing the ground truth hypotheses, there should be no element in a hypothesis that is not present in any ground truth hypothesis.
- Knowing the ground truth hypotheses, there should be no element in some ground hypothesis which is not in one of the hypotheses generated by HCDX.

The first four properties will be computed without any knowledge of ground truth. Violating any of these properties can lead to very poor confidence scores. One very simple metric (shown in Figure 5) is simply the sum of the counts of all the detected violations of the properties.

5.2 User Experience

When considering large hypotheses spaces, any human interacting with the system will need a method to rapidly explore the space. For HCDX, we define a set of metrics that facilitate exploring the interpretations and highlight the strengths of applying MBD to this problem.

Incoherence = $a_0 + a_1 + a_2 + a_3 + a_4 + a_5$

a_0 number of ill-formed hypotheses.

a_1: number of logically incorrect hypotheses.

a_2: number of incorrectly ruled out hypotheses.

a_3: number of incomplete hypotheses.

a_4: number of KEs not in ground truth.

a_5: number of ground truth KEs not in hypothesis.

Figure 5: Semantic coherence metric with higher scores indicating worse hypotheses.

6 Automated Hypotheses Evaluation

The preferences among interpretations are likely to be more complex than we see in Model-Based Diagnosis applications. We expect the following will be important in ranking hypotheses. Furthermore, these metrics define a space of hypotheses in which a user could ask for hypotheses that are further away in the metric space along the pareto frontier. Our approach combines the following pieces of information:

a) Probability of hypotheses: Interpretation of the confidence values from sources as probabilities and combining them described earlier.

b) Coverage over time: HCDX rates hypotheses in which the temporal extent of the hypothesis's KEs closely matches the temporal extend of the scenario model's KEs. That is, if the scenario model includes events from 12 month period and the hypotheses only contains events from a one month period this time mis-match indicates that there is likely information missing from the resulting hypotheses.

c) Coverage over space: HCDX rates hypotheses in which the spatial extent of the KEs of the hypothesis closely matches the spatial extent of the KEs of the scenario model; i.e., if the scenario model includes events from five countries and the hypothesis only uses events from one city, this mismatch indicates that it is likely that there is information missing from the hypothesis.

d) Coverage over sources/source-types: one objective of is to break the siloing of analysts and analytics tools. Hypotheses that include KEs from different sources and media types are therefore preferred over hypotheses from a single source.

e) Hypotheses coverage: For a given scenario model, hypotheses that explain more of the KEs should be trusted more.

f) Scenario model coverage: In model assembly, HCDX creates a model that corresponds to a subset of the KEs in the KB. If the KEs of the assembled model are a small subset of the KEs in the KB, all hypotheses generated by these models should be discounted.

g) Cost of assumptions: HCDX infers additional KEs and relationships. Therefore, as in [5], hypotheses that rely on this information that was not explicitly provided should be discounted.

h) Depth of inference chains: Building on the systematicity principle from Cognitive Science [6], hypotheses that include long causal chains are favored over hypotheses with shallow connections between assumptions and KEs. These deep inference chains are favored because they provide additional opportunities to refute the hypothesis and rely on less assumables.

6.0.1 Directing Future Analysis

In addition to presenting sets of hypotheses, HCDX enables a type of reasoning that can help users differentiate among their hypotheses. In MBD, generating hypotheses is often only the first step of troubleshooting. HCDX can also answer the question: What additional evidence would be required to differentiate among the hypotheses and how can that evidence be gathered at the lowest cost? HCDX can identify that piece of possible evidence that best differentiates among the alternative hypotheses. That piece of evidence can be presented to the user as it pinpoints exactly where possible hypotheses differ in ways that could be distinguished with more evidence gathering. In addition that piece of evidence can be provided to the module that constructs the common semantic interpretation for additional analysis. This approach can be applied whether or not costs and/or probabilities are provided. If no probabilities or costs are provided, HCDX identifies pieces of evidence to focus on simply from the combinatorial structure alone. If probabilities are provided but no costs, HCDX can use a minimum entropy technique to identify useful pieces of evidence. If both probabilities and costs of gathering evidence are included, HCDX can suggest evidence to through multi-step lookahead.

7 Related Work

Beyond the diagnosis community, there are other large projects focused on understanding systems based on observations. The Big C is designed to integrate complex mechanistic models of cancer biology from reading scientific papers [7]. While the microtheory structure supports contradictions, it is not reconfigurable resulting in ordering bias around hypotheses. While the plan, activity, and intent recognition community generate hypotheses from time series observations [8], they focus on the best explanation and require either a domain model or hand labeled examples. The AHEAD system creates structured arguments about given hypotheses regarding asymmetric threats [9]. The HCDX approach described in this paper does not depend on labeled examples or predefined domain specific rules.

8 Summary

This paper has described an initial approach to constructing multiple hypotheses of scenarios using ideas from model-based diagnosis. We believe this approach has great promise over other approaches for forming hypotheses. However, more work is required to flesh out and refine the ideas. A fully functional tool will require modules for data intepretation, compilation and user interaction which is not to focus of this paper. Any progress we can make on helping people become aware of and differentiate amongst alternative interpretations can have great benefit for our society.

References

[1] J. de Kleer and B. C. Williams. Diagnosing multiple faults. *Artificial Intelligence*, 32(1):97–130, April 1987. Also in: *Readings in NonMonotonic Reasoning*, edited by Matthew L. Ginsberg, (Morgan Kaufmann, 1987), 280–297.

[2] J. de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.

[3] R.A. Rossi and Rong Zhou. Parallel collective factorization for modeling large heterogeneous networks. *Social Network Analysis and Mining*, 6(1):67–, 2016.

[4] Matthew Klenk and Ken Forbus. Analogical model formulation for transfer learning in ap physics. *Artificial Intelligence*, 173(18):1615–1638, 2009.

[5] Scott E Friedman, Kenneth D Forbus, and Bruce Sherin. Constructing and revising commonsense science explanations: A metareasoning approach. In *2011 AAAI Fall Symposium Series*, 2011.

[6] Catherine A. Clement and Dedre Gentner. Systematicity as a selection constraint in analogical mapping. *Cognitive Science*, 15(1):89–132, 1991.

[7] Michael J Witbrock, Karen Pittman, Jessica Moszkowicz, Andrew Beck, Dave Schneider, and Douglas B Lenat. Cyc and the big c: Reading that produces and uses hypotheses about complex molecular biology mechanisms. In *AAAI Workshop: Scholarly Big Data*, 2015.

[8] Gita Sukthankar, Christopher Geib, Hung Hai Bui, David Pynadath, and Robert P Goldman. *Plan, activity, and intent recognition: theory and practice*. Newnes, 2014.

[9] William J Murdock, David W Aha, and Leonard A Breslow. Ahead: case-based process model explanation of asymmetric threats. *Navy Center for Applied Research in Artificial Intelligence Technical Note AIC-02-203. Available at http://www. aic. nrl. navy. mil/˜ aha/papers/AIC-02-203. pdf, accessed October*, 2002.