Foundations of model construction in feature-based semantic science

David Poole

Department of Computer Science,
University of British Columbia,
poole@cs.ubc.ca

Abstract. The aim of what semantic science is to have scientific ontologies, data, and hypotheses represented and published in machine understandable forms that enable predictions on new cases. There is much work on developing scientific ontologies and representing scientific data in terms of these ontologies. The next step is to publish hypotheses that can make (probabilistic) predictions on the published data and can be used for prediction on new cases. The published data can be used to evaluate hypotheses. To make a prediction in a particular case, hypotheses are combined to form models. This paper considers feature-based semantic science where the data and new cases are described in terms of features. A prediction for a new case is made by building a model made up of hypotheses that fit together, are consistent with the ontologies used, and are adequate for the case. We give some desiderata for such models, and show how the construction of such models is a form of abduction. We provide a definition for models that satisfies these criteria and prove that it produces a coherent probability distribution over the values of interest.

Introduction

If a knowledge-based system makes a prediction, it is reasonable for someone to ask: what evidence is there for that prediction? The system should be able to provide such evidence. If a knowledge-based system is to believe something, it should believe it based on evidence, as not all beliefs are equally valid. The mechanism that has been developed for constructing and judging knowledge is called *science*. Science determines truth based on empirical evidence: what does all of the available evidence lead us to conclude?

The semantic web (Berners-Lee et al., 2001) is an endeavor to make all of the world's knowledge accessible to computers. One of the central concerns of the semantic web is how to trust the information given. Trust in the truth of some information, or what Gil and Artz (2007) call *content trust*, has been cast in terms of social trust relationships. Search engines such as Google base their ranking on measures such as pagerank (Page et al., 1999) which essentially measure popularity, but these search engines often return authoritative sites. If you are a scientist, popularity and appeal to authority are not the basis for determining what is true.

We use to term *semantic science* (Poole et al., 2008) in an anaolgous way to the *semantic web*, because the computer should understand the hypotheses and data which form the foundation of science. Science is used as the basis for trust; we trust scientific conclusions because they are based on the evidence available.

This paper is about one aspect of semantic science: selecting hypotheses and applying them to new cases. This is reminiscent of abduction (Poole, 1989; Kakas and Denecker, 2002) as proposed by Peirce (Burch, 2008), although we take a probabilistic view, where the predictions are all probabilistic. Note that abduction is usually used for explaining observations, but it can also be used for prediction (Poole, 1989): to determine whether to predict some proposition, we explain why it is (may be) true and explain why it is (may be) false, and then consider which explanations would be more surprising. Probabilistic inference can be cast in these terms (Poole, 1993a). Unlike the normal definition of abduction (Kakas and Denecker, 2002), we treat abduction as building a probabilistic explanation of observations.

Semantic Science Overview

The basic idea of semantic science is:

- Information is published using well defined *ontologies* (Smith, 2003b) to allow semantic interoperability. The ontologies specify the shared vocabulary.
- Observational *data* is published (Fox et al., 2006; McGuinness et al., 2007) using the vocabulary specified in the ontologies. Part of this data includes metadata about what the data is about and how it was generated. Data repositories include the Community Data Portal (`http://cdp.ucar.edu/`) and the Virtual Solar-Terrestrial Observatory (`http://www.vsto.org`).
- Scientists publish *hypotheses* that make predictions on data. These hypotheses make reference to ontologies, so that they can interoperate with each other and with the data. As part of each hypothesis is information about what data this hypothesis is prepared to make predictions about. These predictions can be tested on the published data.
- New data can be used to evaluate, and perhaps update, the hypotheses that make predictions on this data. Predictions on new data can be used to judge the hypotheses as well as find outliers in the data.
- The descriptions of competing hypotheses can be used to devise experiments that will distinguish the hypotheses (see e.g., King et al., 2004).
- If someone wants to make a prediction for a new case (e.g., predicting the outcome of a patient in a diagnostic setting, or predicting landslide susceptibility), multiple hypotheses may need to be combined into a *model* and to be applied to this special case. Note that the hypotheses are general in the sense that they can be applied to multiple cases, but are typically very narrow in that they only make predictions in narrow contexts. The models are constructed for the specific prediction.
- Given a prediction, a user will be able to find out what hypotheses were used for the specific prediction, and then ask for what evidence there is for each hypothesis. In this way, all information will be auditable.
- There is no central authority to vet as to what counts as a legitimate scientific hypothesis. Each of us can choose to make decisions based on whichever criteria we

want. We will be able to judge hypotheses by their predictions on unseen data and other criteria.

- We expect semantic science search engines to be developed. Given a hypothesis, a search engine would be able to find data that can be used to evaluate or tune the hypothesis. Given data, a search engine would be able to find the hypotheses that make predictions on the data.

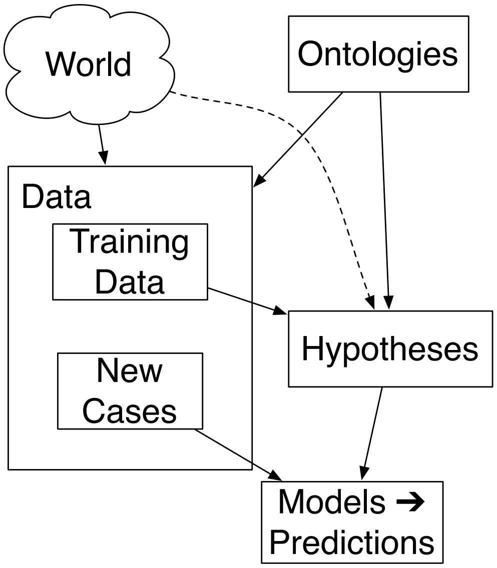


Fig. 1. Role of Ontologies, Data, Hypotheses and Models

The relationship amongst ontologies, data, hypotheses and models is given in Figure 1. The data depend on the world and the ontologies. The hypotheses depend on the ontologies, indirectly on the world (if a human is designing the hypotheses), and directly on some of the data (as we would expect that the best hypotheses would be based on as much data as possible). Given a new case, multiple hypotheses can be used to form a model to make predictions about that case. These predictions can be used for decision making. The ontologies, data sets and hypotheses evolve in time.

The term “science” is meant to be as broad as possible. We can have scientific hypotheses about any natural or artificial phenomenon. The scientific method is applicable to any domain. We could have scientific hypotheses about traditional disciplines such as earth sciences, physics, chemistry, biology, medicine and psychology but we would also imagine hypotheses as diverse as predicting which companies will be most profitable, predicting where the best parties are, predicting who will win football games, or even predicting which celebrities are having affairs. The only criterion is that a scientific hypothesis must put itself at risk by making predictions about observable phenomenon.

Semantic science has no prior prejudice about the source or the inspiration of hypotheses; as long as the hypotheses are prepared to make predictions about unseen data, they can be included. We are not, a priori, excluding religion, astrology, or other areas that make claim to the truth; if they are prepared to make predictions about what will be observed, we can test how well their predictions fit the available data, and use their predictions for new cases.

Semantic science is trying to be broad and bottom-up. It should serve to democratize science in allowing not just the elite to create data and hypotheses. Like scientists themselves, it should be skeptical of all of the information it is presented with.

We anticipate that the most useful hypotheses will make probabilistic predictions, however hypotheses can make diverse forms of predictions, such as definitive predictions, qualitative predictions, fuzzy predictions or probability intervals. Users of hypotheses can choose to adopt hypotheses based on whatever criteria they like, e.g., some combination of fit to the existing data and simplicity or prior plausibility. Users can also choose to ignore hypotheses that don't make the sort of predictions they like.

For the foreseeable future virtually all (useful) hypotheses will be a mix of human generated and machine learned; humans define the structure and parameter space and the machines optimizes these with respect to fit to data and learning biases. This adds new challenges to machine learning: cope with multiple persistent heterogenous data sets that are published with respect to formal ontologies. The hypotheses need to take into account rich meta-data about observations.

Semantic science provides a mechanism for Bayesian inference, where we should condition on *all* relevant information that was not part of building the model. A semantic science search engine should allow us to find all of the relevant data on which to condition.

The hypotheses here are meant to be general hypotheses that can be applied to new cases. Hypotheses about a particular case (e.g., hypotheses about what is wrong with a component for diagnosis) are carried out by the models. The hypotheses in this paper are ones that can make predictions and be evaluated against multiple data sets.

To make this project manageable, we can define four levels of semantic science:

0. Deterministic semantic science where all of the hypotheses make definitive predictions. This class includes both propositional and first-order hypotheses. This has been studied under the umbrella of abductive logic programming (Poole, 1989; Kakas and Denecker, 2002).
1. Feature-based semantic science, with non-deterministic[1] predictions about feature values of data. Learning feature-based representations is the most common form of machine learning. Such hypotheses can be specified in terms of random variables that represent the values of features. We assume that the set of possible feature values are specified as part of the ontology.
2. Relational semantic science, where the predictions are about the properties of individuals and relationships among individuals. In this case, the values of properties may be meaningless names of individuals; the structure of the relationships is used

[1] Non-deterministic can mean many things. Here we consider just the case where there are probabilistic predictions. But there are many alternatives, such as qualitative predictions, probability ranges or fuzzy predictions.

to make predictions. This is what has been studied in inductive logic programming (Muggleton and De Raedt, 1994) and statistical relational learning (Getoor and Taskar, 2007; De Raedt et al., 2008).

3. First-order semantic science, where the aim is to make predictions about the existence of individuals or predictions about universally quantified statements. This is more challenging as conditioning is not well-defined (Poole, 2007). We may not know which individual in the world a hypothesis is making a prediction about, as the hypothesis may refer to the existence of individual filling a role, but we may not observe which individual fills the role.

In the rest of this paper, we consider the second of these where data and hypotheses are described in terms of features. Features generalize propositions, as a proposition can be seen as a Boolean feature. Features can also be seen as properties of a single individual under consideration. There can also be global features that are not about any individual. Treating features as properties allow for a correspondence with the work on ontologies. We use the term *attribute* for a feature-value (or property-value) pair, for example that a rock's age is 50 million years is an attribute of the rock.

Formalizing Feature-based Semantic Science

In this section we give a simple formalization of feature-based semantic science. To keep the description manageable, we ignore interventions, and only include observational data. We include a running example on diagnosis that is not meant to be realistic, but is designed to highlight the issues.

Ontologies

In AI, an ontology (Smith, 2003b; Noy and Hafner, 1997; Gómez-Pérez et al., 2004) is a specification of the meaning of vocabulary used by an information system. Ontologies form the backbone of the Semantic Web (Berners-Lee et al., 2001). There has recently been much work in standardizing ontologies, such as using the Web Ontology Language OWL (Hitzler et al., 2009). Science is one of the areas where ontology development and deployment is well under way (Smith et al., 2007).

Ontologies can be very complicated, as would be expected in a world where language has evolved to be useful and new terminology is invented to describe what was not easy to describe using previous terminology.

We have been advocating a structure for ontologies using what are called Aristotelian definitions (Smith, 2003a; Poole et al., 2009), based on the idea of Aristotle (350 B.C.) that each class should be described in terms of a super-class (the genus) and the attributes (the differentia) that differentiate this class from other subclasses of the genus. Defining all classes in terms of attributes, as opposed to specifying subclass relationships directly, simplifies reasoning as we only need to give the values of properties and the class structure logically follows. It is also a natural way to define concepts in many cases. Simple Aristotelian definitions often give rise to complicated subclass relationships, but simple subclass relationships give simple Aristotelian definitions.

For the rest of this paper, we will thus ignore classes, and consider only features (conflating features and properties as we are only considering feature-based semantic science). Properties (and so features) have domains; they are only defined in the context where other properties have particular values. Properties are not defined when their domain does not hold.

Properties:

Property	Domain	Range
IsPerson	thing	boolean
Age	person	integer
Sex	person	{male,female}
Coughs	person	boolean
HasLump	person	boolean
LumpShape	lump	{circular, oblong, irregular}
LumpLocn	lump	{leg,torso,arm,head}
CancLump	lump	boolean
LumpColour	lump	{red,pink,brown,. . . }
HasCancer	person	boolean
HasLungCancer	personWithCancer	boolean
OutcomeAtYear	person	{well,sick,dead}
TakenH53	person	boolean

Classes:

Class	Genus	Differentia
person	thing	IsPerson=true
lump	person	HasLump=true
personWithCancer	person	HasCancer=true

Fig. 2. Properties and Classes For Running Example

Example 1. Figure 2 shows an ontology used in the running example. Here, *thing* is the top-level class. This can be translated into OWL in a straightforward manner, for example:

```
FunctionalDataProperty(HasLump)
DataPropertyDomain(HasLump person)
DataPropertyRange(HasLump xsd:boolean)
EquivalentClasses(lump
           DataHasValue(HasLump true))
FunctionalDataProperty(CancLump)
DataPropertyDomain(CancLump lump)
DataPropertyRange(CancLump xsd:boolean)
ObjectPropertyDomain(LumpShape lump)
ObjectPropertyRange(LumpShape
   ObjectOneOf(circular oblong irregular))
```

The ontology can also include axioms that specify that people with cancerous lumps have cancer by definition:

```
SubClassOf(DataHasValue(CancLump true)
    personWithCancer)
```

We will use the property *PropertyDomain* to mean OWL's *DataPropertyDomain* or *ObjectPropertyDomain*.

We define a **literal** to be an assignment of a value to a feature. A **proposition** is a formula made of literals and the standard Boolean connectives, with their standard meaning. A **conjunction of literals** is a proposition that only includes the logical-and connective (including the empty conjunction, $true$, and singleton literals). If c is a conjunction of literals, let $features(c)$ be the set of features assigned in c.

With Aristotelean definitions, classes are represented as propositions that define membership in the classes. We assume a top-level class *thing* that corresponds to the proposition $true$.

We write $Ontologies \models w$ to mean proposition w is entailed by the ontologies, for example, in terms of the OWL 2 direct semantics (Motik et al., 2009) or the OWL 2 RDF semantics (Schneider, 2009). For this paper, we assume that the union of all of the ontologies used is satisfiable.

Data

We assume that data about observations of the world are published referring to the ontologies used. For the purpose of this paper[2], assume a **data set** is made up of $\langle c, O, t\rangle$ triples where:

- c is a proposition that specifies the context in which the data was collected.
- O is a set of features that were observed. For this paper, we assume[3] that the context c implies the domain of each feature in O. That is, if for any $o \in O$:

  $$Ontologies \models PropertyDomain(o, d)$$

  then

  $$Ontologies \models c \rightarrow d.$$

- t is a table on O which represents the actual observed values; that is, t is a set of tuples, where each tuple maps each feature $f \in O$ into a value in the range of f.

To predict such data, a hypothesis needs to predict the values of the observed variables given the context.

[2] We ignore other metadata. Metadata includes the provenance of the data (Bose and Frew, 2005), when and where it was collected, what sensors were used, what processing was done on the data, all of which are important for making predictions on the data.

[3] The alternative is to allow undefined values for those features for which the domain doesn't hold due to the value of other features.

Example 2. Suppose we have data about people who came into a doctor's office. One such data set could include:

$< person, \{Age, Sex, Coughs, HasLump\},$

Age	Sex	$Coughs$	$HasLump$
23	$male$	$true$	$true$
...	...	...	...

$\rangle$

We could also have data about those people with lumps:

$< lump,$
$\{LumpLocn, LumpShape, LumpColour, CancLump\},$

$LumpLocn$	$LumpShape$	$LumpColour$	$CancLump$
leg	$oblong$	red	$false$
...	...	...	...

$\rangle$

We could also have data that was only collected for people who have cancer:

$< personWithCancer,$
$\{HasLungCancer, TakenH53, Age, OutcomeAtYear\},$

$HasLungCancer$	$TakenH53$	Age	$OutcomeAtYear$
$true$	$true$	77	$dead$
...	...	...	...

$\rangle$

Hypotheses

Each hypothesis makes predictions about some feature values.

We assume a **hypothesis** is made up of $\langle c, I, O, P \rangle$ tuples consisting of:

- a context c, which is a proposition that specifies when the hypothesis can be applied
- a set I of input features about which it does not make predictions
- a set O of output features about which it can make a prediction (as a function of the input features)
- a program P that predicts a distribution over the output features for each combination of values for the input features.

We assume that the context implies the domains of all of the input and the output features. The programs can be arbitrarily complex and use arbitrary computation to make predictions

If $h = \langle c, I, O, P \rangle$, we say that $c = context(h)$, $I = inputs(h)$, $O = outputs(h)$.

For example, the ideal gas law is a hypothesis that makes predictions about the pressure P, volume V, number of particles n and the temperature in the context of a gas, namely that $PV \propto nT$. It makes predictions that can be judged against data. There are alternative hypotheses that are more accurate for real gasses, e.g., when the pressure is very high, or when the gas molecules are heterogeneous.

Hypotheses are not universally applicable. For example, the ideal gas law is not applicable to rocks or to lung cancer; we can't use a hypothesis about the prognosis of people with cancer on rocks or gasses. Hypotheses have preconditions that specify what they make predictions about. These preconditions are of three different sorts:

- Conditions which define when the hypothesis makes sense. When these conditions are false, the hypothesis is nonsense. The conditions must imply the domains of the features used in the hypothesis.
- Conditions which define the intended scope of the hypothesis. These conditions specify what the hypothesis was designed to predict.
- Conditions which specify when the hypothesis will be used in a particular case.

For example, a hypothesis that makes predictions of the prognosis of patients with lung cancer may be applicable for arbitrary people. In a particular model, it may only be used for the patients with lung cancer who have not a eaten particular herb ($H53$, below), as the model may use another hypothesis that makes predictions in that case.

One class of hypotheses that is of particular interest is the "null hypothesis". There is a null hypothesis for each feature[4]. This hypothesis says that the feature has randomly distributed values, with probabilities that are independent of the other features. It is important as it is always applicable, and gives a base case upon which to compare other hypotheses.

Example 3. Consider the following hypotheses:

- H_1 predicts the prognosis of people with lung cancer:

$$\langle personWithCancer \wedge HasLungCancer = true,\\ \{\}, \{OutcomeAtYear\}, P_1\rangle$$

- H_2 predicts the prognosis of people with cancer:

$$\langle personWithCancer, \{\}, \{OutcomeAtYear\}, P_2\rangle$$

- H_3 is a null hypothesis that predicts the prognosis of people in general:

$$\langle person, \{\}, \{OutcomeAtYear\}, P_3\rangle$$

- H_4 predicts whether people with cancer have lung cancer, as a function of coughing:

$$\langle personWithCancer, \{Coughs\},\\ \{HasLungCancer\}, P_4\rangle$$

- H_5 predicts whether people have cancer:

$$\langle person, \{\}, \{HasCancer\}, P_5\rangle$$

- H_6 and H_7 predict the shape of lumps as a function of whether the patient has cancer[5]. We first predict whether the patient has a lump, using H_6:

$$\langle person, \{HasCancer\}, \{HasLump\}, P_6\rangle$$

  and then predict the lump shape when there is a lump using H_7:

$$\langle lump, \{HasCancer\}, \{LumpShape\}, P_7\rangle$$

[4] There are actually infinitely many null hypotheses, one for each probability distribution, but we usually only consider the maximum likelihood or maximum apriori probability null hypothesis.

[5] If we were to allow null values, we could replace H_6 and H_7 with H_8:

$$\langle person, \{HasCancer\}, \{HasLump, LumpShape\}, P_8\rangle$$

where we would make sure that *LumpShape* is undefined when $HasLump = false$.

- H_9 predicts coughing of people

  $$\langle person, \{\}, \{Coughs\}, P_9 \rangle$$

- H_{10} predicts coughing of people as a function of whether they will live a year:

  $$\langle person, \{OutcomeAtYear\}, \{Coughs\}, P_{10} \rangle$$

  While this may seem like a peculiar hypothesis, it may be the sort of hypothesis one would get from a study that selected 1000 people who died after a year and 1000 people who didn't die, and then checked whether they had reported coughing.
- H_{11} predicts whether the individual under consideration is a person:

  $$\langle thing, \{\}, \{person\}, P_{11} \rangle$$

A hypothesis $\langle c, I, O, P \rangle$ is a representation of the conditional distribution:

$$P(O|I, c).$$

That is, P gives a probability distribution over O as a function of I for context c.

Note that, if a hypothesis makes a prediction on features O, in principle, it can be used to make predictions on subsets of O. However, it is not always computationally feasible to sum out the variables needed to compute this.

Models and Predictions

Scientific hypotheses are typically narrow; they don't make predictions on arbitrary sets of data. For example, someone may develop a hypothesis for the prognosis of a particular type of lung cancer. To use this hypothesis for a prediction of a patient, we first predict whether the patient has this form of lung cancer, then use this hypothesis to predict the prognosis. We need other hypotheses about the prognosis for the possibility that the patient has a different form of lung cancer, or doesn't have lung cancer.

We assume that a new case includes observations and a set of query features that we want to predict the value of. In particular, a **query** is a pair $\langle obs, Q \rangle$ where obs is a conjunction of literals and Q is a set of features. This query is asking for a prediction of $P(Q|obs)$, the distribution of Q given observations obs. We assume that obs does not assign a value to a feature in Q.

A set of hypotheses that fit together to make a prediction for the query variables given the observations is called a *model*. Before giving a definition of a model, we will give some desiderata of models. Note that we use the term *model* here in the sense of scientific models, not in the sense used in logic.

A model M for query $\langle obs, Q \rangle$ needs to satisfy the following properties:

- M is coherent: it does not rely on the value of a feature in a context where the features is not defined (i.e., when the domain of the feature is false). Thus if feature f has domain d, it has to be used in a context where d is true. For example, always writing $d \wedge f$, which is false if d is false, and has the value of f otherwise, would satisfy coherence.

- M is consistent: it does not make different predictions for any feature in any particular context.
- M is predictive: it makes a prediction for Q in every context that is possible given the observations.
- M is minimal: it does not include hypotheses that are not required to be predictive.

For level-0 (deterministic) semantic science, these desiderata correspond to a standard definition of abduction (Poole, 1989; Kakas and Denecker, 2002). Coherence is needed when there are ontologies with non-trivial domains of properties. The predictive condition corresponds to being able to prove the goal, and the consistency and minimality are the same as in the standard definition of abduction.

For level-1 (feature-based) semantic science, ignoring preconditions, one way to build a model is to construct a Bayesian network (Pearl, 1988). The variables in the Bayesian network correspond to features. For every variable in the Bayes network, there is a corresponding hypothesis with that variable in the set of output features of the hypothesis. Such a Bayesian network needs to include as variables: the observation variables, the query variables and the inputs for every hypothesis used. The accuracy of such a Bayesian network depends on the accuracy of the hypotheses used as well as the appropriateness of the independence assumptions embedded in the Bayesian network.

When hypotheses have preconditions, we need to ensure that the preconditions hold before being able to use the hypotheses. We want to be able to use some hypotheses in some contexts and not in other contexts. We do not need a global acyclic assumption, but we disallow cyclic dependencies. This leads to the following definitions, where we first define hypothesis instances, which are the building blocks of models.

A **hypothesis instance** is a tuple of the form $\langle h, c, I, O \rangle$ such that:

- h is a hypothesis,
- c is a a conjunction of literals, which specifies a context in which hypothesis h will be used
- I is a set of input properties used by hypothesis h
- O is a set of output properties which hypothesis h will be used to predict

satisfying the following:

- $Ontologies \models c \rightarrow context(h)$ — the condition in which the hypothesis is used must imply the context of the hypothesis h, and so it must imply the domains of the features used in h
- $inputs(h) \subseteq I \cup features(c)$ — the inputs to the hypothesis must all be available, either in c or in I
- $O \subseteq outputs(h)$ — not all of the outputs need to be used

A model is a set of hypothesis instances that together define a probability distribution of a query given observations. To formalize this, satisfying the desiderata above, and being as general as possible, requires a syntactic construction that allows for different orderings and different features to be defined in different contexts. If a hypothesis instance is applicable in a context it has to be used in that context. This motivates the following definitions.

A set M of hypothesis instances is **structurally consistent** if for every pair of different hypothesis instances $\langle h_1, c_1, I_1, O_1 \rangle$ and $\langle h_2, c_2, I_2, O_2 \rangle$ in M, if $O_1 \cap O_2 \neq \{\}$, then $ontologies \models \neg(c_1 \wedge c_2)$. That is, if they make predictions on the same feature, their contexts must be incompatible. This is called structural consistency as it only takes into account the structure of the hypothesis instances, and not on the details of the actual prediction made.

A **semantic tree**[6] is a tree where:

- internal nodes are labelled with features
- there are children of the node for each value of the feature
- a feature appears at most once in any path from the root.

Each path from the root corresponds to a set of feature-value pairs, which we interpret as a proposition made from the conjunction of the corresponding literals. The root corresponds to the proposition $true$.

Given a set M of hypothesis instances, a semantic tree **built from** M must satisfy the following conditions:

- A node is labelled with feature f only if there is a $\langle h, c, I, O \rangle \in M$ with $f \in O$ such that c is entailed by the path to the node and every feature in I appears as the label of an ancestor of the node. In this case, $\langle h, c, I, O \rangle$ is the hypothesis instance for the node.
- If $\langle h, c, I, O \rangle$ is a hypothesis instance for a node then, for every path from the root that goes through the node, there must be a node labelled with each element of O.

A semantic tree supports **supports** query $\langle obs, Q \rangle$ if every path from the root either implies $\neg obs$ or implies obs and every element of Q appears in the path.

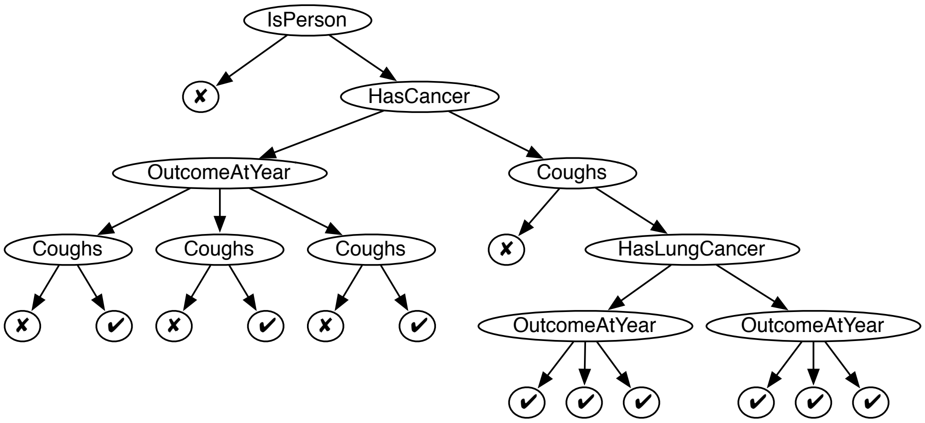
A **model** M for query $\langle obs, Q \rangle$ is a structurally consistent set of hypothesis instances for which there is a semantic tree built from M that supports $\langle obs, Q \rangle$.

Example 4. Suppose we have data about a person who coughs and has an irregular lump, and we want to make predictions about their outcome in a year.

A possible model for $P(OutcomeAtYear | person \wedge coughs)$:

- $\langle H_5$, $person$, $\{\}$, $\{HasCancer\}\rangle$
- $\langle H_3$, $person \wedge \neg hasCancer$, $\{\}$, $\{OutcomeAtYear\}\rangle$
- $\langle H_9$, $personWithCancer$, $\{\}$, $\{Coughs\}\rangle$
- $\langle H_4$, $personWithCancer$, $\{Coughs\}$, $\{HasLungCancer\}\rangle$
- $\langle H_1$, $personWithCancer \wedge hasLungCancer$, $\{\}$, $\{OutcomeAtYear\}\rangle$
- $\langle H_2$, $personWithCancer \wedge \neg hasLungCancer$, $\{\}$, $\{OutcomeAtYear\}\rangle$
- $\langle H_{10}$, $person \wedge \neg hasCancer$, $\{OutcomeAtYear\}$, $\{Coughs\}\rangle$
- $\langle H_{11}$, $thing$, $\{\}$, $\{IsPerson\}\rangle$

[6] There are many names for such trees, including event trees, decision trees, game trees, or computation trees (Halpern, 2003), each of which seems to convey a different intuition. There are no events or time involved in these trees, they just represent assignments of values to features.



Left branches from a node correspond to $false$ and right branches to $true$, except for nodes labelled with $OutcomeAtYear$, where the three children correspond to the values $well$, $sick$, $dead$. The leaves are labelled by whether the observations are true or not.

Fig. 3. A semantic tree for Example 4

In this model, although H_2 can make predictions for anyone with cancer, it is only used for those without lung cancer. Similarly H_3 is only used when the person does not have cancer.

A semantic tree built from M is shown in Figure 3.

Not all features are defined in all contexts. In particular, *HasLungCancer* is not defined in the context $person \wedge \neg hasCancer$.

There is no global ordering of the features; *OutcomeAtYear* needs to be above *Coughs* in one context and below it in another context.

Semantics of Models

A semantic tree built from a model for $\langle obs, Q\rangle$ provides a possible world structure, that defines $P(Q|obs)$ using only conditional probabilities from the hypotheses of the model in the context defined by the model.

As we are assuming finitely many features with finite domains, there are only finitely many possible worlds, so it suffices to give a probability to each world. The possible world structure is complicated because the worlds are heterogeneous; not all features have values in all worlds. The proposition *obs* will be well defined in all worlds, and in all worlds where *obs* is true, Q will be defined. The possible world structure is built from a semantic tree, which in turn depends on the model. Different models can give different possible world structures, as different features can be defined. We show that the distribution for $P(Q|obs)$ for a particular model does not depend on which semantic tree is built from the model.

The worlds for a semantic tree correspond to paths from the root in the tree. The feature-value pairs along the path are all true in the world, the other values for the features along the path are all false, and all other features are undefined in the world. Note that our semantics does not include an $undefined$ value; we never need to use the values of an undefined feature.

The probability of a world is the product of the probabilities computed by the programs of the hypotheses used by the model that are consistent with the world. That is, the probability of world w for model M is the product of the numbers $P(O=o|I=i,c)$ such that $\langle h,c,I,O\rangle \in m$, the world w entails $O=o, I=i, c$, and P is the program of h.

Lemma 1 (coherence). *What is true in any world can be determined without reference to any feature that is not defined in the world.*

Proof. A hypothesis is only used in a world if its context is true in that world. As its context is true, this implies the domains of the features used in the hypothesis are true. As all of the features are defined through hypotheses, for each features that is given a value in a world, its domain holds in the world.

Lemma 2 (consistency). *A world gives a consistent prediction for each feature it defines.*

Proof. On each path from the root, by definition, a feature appears at most once, and so it has a unique value. Those features that do not appear in the path are not defined in the world.

Lemma 3. *A tree gives a probability distribution over worlds.*

Proof. We need to show that the probabilities of the words are non-negative and sum to 1. They are all non-negative as they are the product of non-negative numbers. Thus we need to show that the probabilities of the worlds sum to 1.

First consider the case where there are only single hypotheses in the output sets of hypothesis instances. For this case, the lemma can be proved by induction over the size of the tree. If the tree just contains the root, there is one possible world (with no features defined), with probability 1. Suppose the lemma is true for trees with $n \geq 0$ internal nodes, then a tree with $n+1$ internal nodes has (at least) one internal node with only leaf children. The probability of the children of that node sum to the probability of that node (as the hypothesis used gives a distribution over its children), which is the probability of the world with that internal node as a child. So by induction, the sum of the probabilities of all worlds sums to 1.

Where there can be multiple hypotheses in the output sets, we can treat the multiple hypotheses as a composite random variable, and the lemma holds. Splitting on the variables in turn does not change the distribution, neither does moving the splits up and down the tree.

A **maximal semantic tree** for model M is a semantic tree that cannot be extended and still be a semantic tree.

Lemma 4. *Given model M, all maximal semantic trees built from M have the same set of possible worlds, with the same probabilities.*

Proof. Suppose T_1 and T_2 are two maximal semantic tree. For every world w_1 from T_1 there must be a world w_2 from T_2 that is consistent with w_1; as the worlds in T_2 cover all possible cases (intuitively w_1 can be filtered down T_2, where for every feature defined in w_1, world w_1 goes down the appropriate branch and for every feature not defined in w_1 it goes to an arbitrary child). If there is some feature that is defined in w_1 or w_2 that is not defined in the other, it can be consistently added to that world thus contradicting maximality.

The probabilities of w_1 and w_2 do not depend on the trees, and so must be the same.

Theorem 1 (predictiveness). *All semantic trees for a model for query $\langle obs, Q\rangle$ give the same probability distribution for $P(Q|obs)$.*

Proof. A semantic tree T can be extended into a maximal semantic tree T' by doing all possible splits. By the previous lemma, T' gives a unique probability distribution for $(Q|obs)$, that does not depend on T. It is easy to see that by summing out the variables introduced, T has the same distribution as T'.

Note that, as all of the semantic trees give the same distribution, we may as well use the minimal models and the minimal semantic trees.

This construction makes implicit independence assumptions, namely that, for each context, a variable is independent of its non-descendants, given the input of the hypotheses used for that variable in that context. It is thus making an assumption of context-specific independence (Boutilier et al., 1996).

Comparison with other proposals

We can view the semantic tree based semantics as a form of abduction, where the paths from the root correspond to explanations. The probability of each explanation is the product of the probabilities given by the hypotheses used. This use of abduction to compute conditional probabilities was first done by Poole (1993b) in probabilistic Horn abduction. This work generalizes that work by allowing for more general specifications of conditional probabilities (P in a hypotheses does not need to be specified by a logic program), allowing for multiple incompatible hypotheses, and allowing the interaction with ontologies, but is restricted to be feature-based, whereas Poole (1993b) defined a probabilistic relational language.

This work is also closely related to work on dynamic construction of Bayesian networks (Horsch and Poole, 1990; Breese, 1992; Wellman et al., 1992; Laskey, 2008). The other proposals do not allow for multiple competing hypotheses, but are rather designed to be flexible ways to define large relational models, where the models can be defined before the individuals (and thus the random variables) are known. We also extend the dynamic construction work to include ontologies.

Conclusion

This paper has defined models that can be built from hypotheses to give predictions in particular cases. This is the first step in bringing the vision of semantic science to reality.

For feature-based semantic science, the main open problems are how to find the appropriate hypotheses for a query, assemble them efficiently, determine when one model is better than another for a particular query, and judge the quality of hypotheses. Not only must the conditional probabilities be accurate, but the independence assumptions embedded in the model need to be appropriate. We also need to develop relational and first-order versions of the theory.

The potential of semantic science seems huge, but there are many technical and social issues that need to be solved before it can become reality. The development of ontologies and the publishing of data using those ontologies has advanced greatly in recent years. The main technical issues remaining are to do with the representations of the hypotheses and models and the infrastructure to publish and search for data and hypotheses. To bring this vision of semantic science to fruition will require advances in many fields.

Bibliography

Aristotle (350 B.C.). *Categories*. Translated by E. M. Edghill, `http://www.classicallibrary.org/Aristotle/categories/`.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, May: 28–37. URL `http://www.sciam.com/article.cfm?id=the-semantic-web`.

Bose, R. and Frew, J. (2005). Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys*, 37(1): 1 – 28. URL `http://portal.acm.org/citation.cfm?id=1057978`.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). Context-specific independence in Bayesian networks. In E. Horvitz and F. Jensen (Eds.), *UAI-96*, pp. 115–123. Portland, OR.

Breese, J.S. (1992). Construction of belief and decision networks. *Computational Intelligence*, 8(4): 624–647.

Burch, R. (2008). Charles Sanders Peirce. *The Stanford Encyclopedia of Philosophy*. URL `http://plato.stanford.edu/archives/spr2008/entries/peirce/`.

De Raedt, L., Frasconi, P., Kersting, K., and Muggleton, S.H. (Eds.) (2008). *Probabilistic Inductive Logic Programming*. Springer.

Fox, P., McGuinness, D., Middleton, D., Cinquini, L., Darnell, J., Garcia, J., West, P., Benedict, J., and Solomon, S. (2006). Semantically-enabled large-scale science data repositories. In *5th International Semantic Web Conference (ISWC06)*, volume 4273 of *Lecture Notes in Computer Science*, pp. 792–805. Springer-Verlag. URL `http://www.ksl.stanford.edu/KSL_Abstracts/KSL-06-19.html`.

Getoor, L. and Taskar, B. (Eds.) (2007). *Introduction to Statistical Relational Learning*. MIT Press, Cambridge, MA.

Gil, Y. and Artz, D. (2007). Towards content trust of web resources. *Journal of Web Semantics*, 5(4): 227–239. doi:doi:10.1016/j.websem.2007.09.005. URL `http://dx.doi.org/10.1016/j.websem.2007.09.005`.

Gómez-Pérez, A., Fernández-López, M., and Corchu, O. (2004). *Ontological Engineering*. Springer.

Halpern, J.Y. (2003). *Reasoning about Uncertainty*. MIT Press, Cambridge, MA.

Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., and Rudolph, S. (2009). *OWL 2 Web Ontology Language Primer*. W3C. URL `http://www.w3.org/TR/owl2-primer/`.

Horsch, M. and Poole, D. (1990). A dynamic approach to probabilistic inference using Bayesian networks. In *Proc. Sixth Conference on Uncertainty in AI*, pp. 155–161. Boston.

Kakas, A. and Denecker, M. (2002). Abduction in logic programming. In A. Kakas and F. Sadri (Eds.), *Computational Logic: Logic Programming and Beyond*, number 2407 in LNAI, pp. 402–436. Springer Verlag. URL `http://www2.cs.kuleuven.be/cgi-bin/dtai/publ_info.pl?id=39495`.

King, R., Whelan, K., Jones, F., Reiser, P., Bryant, C., Muggleton, S., Kell, D., and Oliver, S. (2004). Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427: 247–252. URL http://www.doc.ic.ac.uk/~shm/Papers/Oliver_Jan15_hi.pdf.

Laskey, K.B. (2008). MEBN: A language for first-order Bayesian knowledge bases. *Artificial Intelligence*, 172(2-3): 140–178. doi:10.1016/j.artint.2007.09.006.

McGuinness, D., Fox, P., Cinquini, L., West, P., Garcia, J., Benedict, J.L., and Middleton, D. (2007). The virtual solar-terrestrial observatory: A deployed semantic web application case study for scientific research. In *Proceedings of the Nineteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-07)*. Vancouver, BC, Canada. URL http://www.ksl.stanford.edu/KSL_Abstracts/KSL-07-01.html.

Motik, B., Patel-Schneider, P.F., and Grau, B.C. (Eds.) (2009). *OWL 2 Web Ontology Language Direct Semantics*, volume W3C Recommendation 27 October 2009. W3C. URL http://www.w3.org/TR/owl-direct-semantics.

Muggleton, S. and De Raedt, L. (1994). Inductive logic programming: Theory and methods. *Journal of Logic Programming*, 19,20: 629–679.

Noy, N.F. and Hafner, C.D. (1997). The state of the art in ontology design: A survey and comparative review. *AI Magazine*, 18(3): 53–74. URL http://www.aaai.org/Library/Magazine/vol18.php\#Fall.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford InfoLab. URL http://dbpubs.stanford.edu/pub/1999-66.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Poole, D. (1989). Explanation and prediction: An architecture for default and abductive reasoning. *Computational Intelligence*, 5(2): 97–110.

Poole, D. (1993a). Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, 64(1): 81–129.

Poole, D. (1993b). Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, 64(1): 81–129.

Poole, D. (2007). Logical generative models for probabilistic reasoning about existence, roles and identity. In *22nd AAAI Conference on AI (AAAI-07)*. URL http://cs.ubc.ca/~poole/papers/AAAI07-Poole.pdf.

Poole, D., Smyth, C., and Sharma, R. (2008). Semantic science: Ontologies, data and probabilistic theories. In P.C. da Costa, C. d'Amato, N. Fanizzi, K.B. Laskey, K. Laskey, T. Lukasiewicz, M. Nickles, and M. Pool (Eds.), *Uncertainty Reasoning for the Semantic Web I*, LNAI/LNCS. Springer. URL http://cs.ubc.ca/~poole/papers/SemSciChapter2008.pdf.

Poole, D., Smyth, C., and Sharma, R. (2009). Ontology design for scientific theories that make probabilistic predictions. *IEEE Intelligent Systems*, 24(1): 27–36. URL http://www2.computer.org/portal/web/computingnow/2009/0209/x1poo.

Schneider, M. (Ed.) (2009). *OWL 2 Web Ontology Language: RDF-Based Semantics*. W3C Recommendation, 27 October 2009. URL http://www.w3.org/TR/owl2-rdf-based-semantics/.

Smith, B. (2003a). The logic of biological classification and the foundations of biomedical ontology. In D. Westerståhl (Ed.), *Invited Papers from the 10th International Conference in Logic Methodology and Philosophy of Science*. Elsevier-North-Holland, Oviedo, Spain. URL http://ontology.buffalo.edu/bio/logic_of_classes.pdf.

Smith, B. (2003b). Ontology. In L. Floridi (Ed.), *Blackwell Guide to the Philosophy of Computing and Information*, pp. 155—166. Oxford: Blackwell. URL http://ontology.buffalo.edu/smith/articles/ontology_pic.pdf.

Smith, B. et al. (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11): 1251–1255. URL http://www.nature.com/nbt/journal/v25/n11/pdf/nbt1346.pdf.

Wellman, M., Breese, J., and Goldman, P. (1992). From knowledge bases to decision models. *Knowledge Engineering Review*, 7(1): 35–53.