What's New?
A Semantic Definition of Novelty

Russell Greiner and Michael R. Genesereth

Heuristic Programming Project

Computer Science Department

Stanford University

Abstract

A central process in any learning experience is the incorporation of a new fact into an existing theory. Despite the abundance of papers on learning, no one has yet defined rigorously what it means to be "new". This paper attempts to fill that gap by first stating, and then formalizing several intuitive ideas about novelty, focusing on what it means for a statement to be a new fact about some concept. The report also includes a brief discussion of how this result might be applied and outlines many remaining research areas.

1 Introduction

A central process in any learning experience is the incorporation of a *new* fact into an existing theory. Often the goal of that process is more specific, to learn some new fact about some concept. But what does it mean to claim that a sentence is *new*, and even more interesting, what qualifies as a novel fact about some concept? Despite the vast interest in learning and the abundance of related papers (cf. [Dietterich 81a], [Buchanan 78], [Michalski 83], [Dietterich 81b], [Dietterich 82]), no one has rigorously defined what it means to be "new", either in general or with respect to a single concept.

This paper attempts to fill that gap. Our goal is to obtain a *semantic* rather than a *syntactic* understanding of novelty. This preference stems from our belief that a semantic account (one based on the possible interpretations of the theory) provides important insight into the phenomenon of novelty. It also means we may be able to generalize these results to other logics and languages.

The results of this research are relevant (and useful) to many different fields. The primary importance of this work is in providing a first stab at describing the different senses of novelty. In addition to the applications a complete and adequate definition of newness would have as an analytic tool, there are possible applications in knowledge acquisition, representation, and discourse analysis. Many of these stem from the intimate connection between novelty and the intuitive notion of "aboutness". (Section 4 elaborates on each of these.)

This report discusses two kinds of novelty. Section 2 describes newness of a sentence with respect to a theory. Section 3 uses this result to address the more difficult task of determining when a sentence conveys something new about a particular concept. While the first kind of novelty is fairly easy to capture, the second requires a consideration of the interconnections among facts within a theory. Section 4 justifies why this undertaking is relevant and describes how these results may eventually be used. The concluding Section 5 lists several outstanding research issues.

2 New with respect to a Theory

This section addresses the issue of what it means for a sentence σ to be new with respect to a theory[1] **Th**; this is the relation $\mathbf{N}(\mathbf{Th}, \sigma)$. Intuitively, we want σ to be new if it (somehow) further specifies something about the world. Alternatively we can think of a new sentence as providing some additional constraints, which remove some possible worlds ([Moore 80]) from consideration.

We first consider a semantic definition of newness: σ is new with respect to a theory if it eliminates some possible interpretation of that theory.[2] That is, given any theory **Th**, in the language **L**, there is a set of models $\mathbf{I}^{\mathbf{Th}} = \{\mathbf{I}_j\}$, in which each $\mathbf{I}_j$ maps the symbols of **L** into objects or sets of tuples of objects in the "real world" in the standard way. Notice that this means that the universe is fixed beforehand and that these ranges can overlap.

Adding additional sentences to a theory can only restrict the set of possible interpretations: $\mathbf{Th} \subseteq \mathbf{Th}'$ means that $\mathbf{I}^{\mathbf{Th}'} \subseteq \mathbf{I}^{\mathbf{Th}}$.

---

[1] We are taking a slightly unorthodox *syntactic* view of theory: viz., a theory is a consistent and deductively closed set of axioms. We will also assume that the deductive system is complete.

[2] We are only concerned with "true interpretations", which map symbols into referents within a *legal* model. We chose "interpretation" rather than "model" to emphasize that we are dealing with a mapping rather than its range.

This leads to the proposal that

Defn 1: $N_{Sem}(Th,\sigma) \Leftrightarrow I^{Th+\sigma} \subsetneq I^{Th}$.

This same definition can be expressed syntactically (in terms of logical deducibility rather than semantic validity) using

Defn 2: $N_{Syn}(Th,\sigma) \Leftrightarrow (\sigma \notin Th) \wedge (\neg\sigma \notin Th)$.

As these are equivalent (whenever the language of the theory remains fixed; see [Enderton 72],) we will simply use **N**.

3 New with respect to a Concept

In many situations it is not enough to realize that an assertion is new; rather, one often wants to claim that it is a new fact *about some concept*. With this in mind, we define the ternary relation **New(Th,s,σ)** to mean that the assertion σ expresses a new fact about the concept **s** with respect to the theory **Th**. The "learning step" involves adding this sentence σ (along with all of its deductive consequences) to the theory.

What should go into a definition of **New(Th,s,σ)**? Clearly, a necessary condition is that σ be a new fact with respect to the entire theory; that is, **N(Th,σ)**. But beyond that, we want to capture the sense in which σ further specifies the concept **s**, or enables the derivation of additional relevant conclusions about **s**.

This section will present a definition of **New** by proposing a series of "increasingly more nearly correct" descriptions. For simplicity, the examples are taken from propositional logic.

Conjecture 1: Syntactic Method. Most statements which relay information about some concept will contain the symbol that refers to that concept. This leads to the proposed syntactic solution: The sentence σ conveys new information about the symbol **A** if the token "A" is lexically included in the string of tokens which form σ, denoted with the assertion **LexIn("A",σ)** — e.g., **LexIn("A","A∧B")**. For the reasons mentioned above, we will further insist that **N(Th,σ)**. Formally,

Defn 3: $New_{Syn}(Th,s,\sigma) \Leftrightarrow N(Th,\sigma) \wedge LexIn(s,\sigma)$.

Unfortunately, this syntactic condition is neither necessary nor sufficient. To see that it is not necessary, realize that we want **New({A⇔B},"A","B")** to be true, since asserting **B** in this situation means that **A** must now be true, which had not been the case before that assertion.

To show insufficiency is a little trickier. Should **New({A∨B},"A","A⇒B")** be true? We argue the answer is no: In this context, asserting **A⇒B** is the same as asserting **B**, which we know says nothing new about **A**. We clearly need a more powerful method for specifying novelty.

Conjecture 2: Fewer Interpretations. Using the notion of possible interpretations discussed in Section 2, we can define the "interpretation range" of a particular symbol. Let the term $I_j(s)$ designate the "real world" referent of the symbol **s**, given by the interpretation I_j — here, either $\underline{T}$ or $\underline{F}$.[3] We use this to define the interpretation range of the symbol **s**, $I^{Th}(s)$, by

Defn 4: $I^{Th}(s) = \{I_j(s) \mid I_j \in I^{Th}\}$.

As additional facts can only further restrict the range of possible interpretations for any symbol, we have $I^{Th'}(s) \subseteq I^{Th}(s)$ whenever **Th** ⊆ **Th'**. This leads to our second conjecture,

Defn 5: $New_{FI}(Th,s,\sigma) \Leftrightarrow I^{Th+\sigma}(s) \subsetneq I^{Th}(s)$.

This New_{FI} definition seems, at first, adequate. In addition to paralleling the **N** situation, it also resonates nicely with the ideas of Shannon's Information Theory, in which information is tied to the reduction of uncertainty in the distribution of possible values of a signal. (See [Gallager 78].) It also handles the two cases used above to discredit New_{Syn}.

Unfortunately, this New_{FI} requirement does not include all desired cases. There are some sentences that do convey new information in the informal sense outlined above but that do not satisfy this constraint: Start with an empty theory, **Th1** ← {},[4] in the language **L** = {**A**,**B**}. The four possible interpretations are shown in Figure 1. By inspection, I^{Th1}("A") = {$\underline{T}$,$\underline{F}$}. Now, form **Th2** ← **Th1** + "A⇔B". While this leaves only two of the four original interpretations, I_0 and I_3, I^{Th2}("A") remains {$\underline{T}$,$\underline{F}$}, indicating that **A⇔B** said nothing New_{FI} about **A**.

Although New_{FI} rejects this **A⇔B**, we still believe it should be considered a new fact about **A** in this situation: If we later learn ¬**B**, we will be able to infer that ¬**A**, a conclusion that would not

	A	B
I_0	$\underline{F}$	$\underline{F}$
I_1	$\underline{F}$	$\underline{T}$
I_2	$\underline{T}$	$\underline{F}$
I_3	$\underline{T}$	$\underline{T}$

Figure 1: Interpretations of A and B in Th1.

have followed without that sentence. That is, any sentence that makes **A**'s range of interpretations dependent on some other symbol (in the sense that **A⇔B** made **A** dependent on **B**) also "feels" new.

So there are at least two ways a statement can be new:

- It directly limits the interpretation range of **A**, or
- It establishes (or increases) a dependency of **A** on some other symbols (as that may, in turn, lead to a reduction of the above type).

While New_{FI} covers the first case exactly, it fails in the second.

Conjecture 3: Partial Interpretations. To define dependency requires an understanding of what it means for one symbol to depend on some other symbols. The **A⇔B** case above is clearly one instance of this. In addition to such singular dependencies (of **A** on one other symbol,) **A** may depend on a combination of symbols. (Consider the assertion **A⇔(B⇔C)**.

[3] The underbar notation denotes the referent of the corresponding linguistic symbol.

[4] The notation **T** ← ψ means the theory **T** is assigned the deductive closure of the set, ψ; and **Th** + σ refers to the deductive closure of **Th**∪{σ}.

Fixing any assignment to **B** alone, **A** will still be "arbitrary"; that is, it could be either $\underline{T}$ or $\underline{F}$, depending on **C**. However, if both **B** and **C** are fixed, then **A** is fully determined.)

We saw that σ is a new fact about **A** if it increases **A**'s dependency on some n-tuple of symbols $\langle s_1, \ldots, s_N \rangle$, that is, if asserting σ restricts the set of assignments available to **A**, given some assignment $\langle \underline{S}_1, \ldots, \underline{S}_N \rangle$ to the symbols $\langle s_1, \ldots, s_N \rangle$. For example, we noted that **A** was more dependent on **B** in **Th2** than in **Th1**. We see this by considering the assignment of **B** to $\underline{F}$. In **Th1**, **A**'s value could be either $\underline{F}$ or $\underline{T}$, independent of this assignment to **B**. However, **A** can no longer be assigned $\underline{T}$ in **Th2**, given this assignment to **B**; its value is now restricted to $\underline{F}$.

As this "$\langle s_1, \ldots, s_N \rangle$ assignment to $\langle \underline{S}_1, \ldots, \underline{S}_N \rangle$" reflects an assignment of only a subset of the symbols, $\{s_i\} \subseteq$ **L**, we will call it a *partial interpretation*. We can associate with each partial interpretation the equivalent class of full interpretations that agree on the assignment of a set of symbols. Formally, take any function that maps some of the symbols of the language into the universe, **U** — that is, any $\varphi : \xi \mapsto \mathbf{U}$, where $\xi \subseteq$ **L**. We can use this to define the equivalence class $[\![\varphi^{\mathbf{Th}}]\!]$ as the set of interpretations that consistently extend φ — that is, it includes each interpretation that agrees with φ's assignment of each symbol in φ's domain and assigns every other element of **L** in some consistent manner.

Defn 6: $[\![\varphi^{\mathrm{Th}}]\!] = \{ I \in \mathbf{I}^{\mathrm{Th}} \mid \forall x \in \mathrm{Domain}[\varphi].\ \varphi(x) = I(x) \}$.

The assignments of **s** that are consistent with the partial interpretation $[\![\varphi^{\mathbf{Th}}]\!]$ are just

Defn 7: $[\![\varphi^{\mathrm{Th}}]\!](s) = \{ I(s) \mid I \in [\![\varphi^{\mathrm{Th}}]\!] \}$.

With this notation, we can state that **A**'s dependency on the symbols $\xi = \mathbf{Domain}[\varphi]$ increases if the set of possible values of **A** consistent with the partial interpretation, $[\![\varphi^{\mathbf{Th}}]\!](\text{"A"})$, decreases but remains non-empty. (Seeing it vanish means that there are no values of **A** that are consistent with this assignment, which means that no models can be derived by extending this partial interpretation.)

To test if σ is new, therefore, consider all of the assignments available to **A** for each partial interpretation, before and after adding this purportedly new sentence σ to the theory. If the number of possible referents of **A** decreases for any partial interpretation (and remains non-empty), we will declare that σ is new.

Defn 8: $\mathrm{New}_{\mathrm{PI}}(\mathrm{Th}, s, \sigma) \Leftrightarrow$
$$\exists \varphi.\ [\![\varphi^{\mathrm{Th}+\sigma}]\!](s) \subsetneq [\![\varphi^{\mathrm{Th}}]\!](s) \wedge [\![\varphi^{\mathrm{Th}+\sigma}]\!](s) \neq \{\}$$

A few notes:

1. We will say that the particular function φ whose partial interpretation $[\![\varphi^{\mathbf{Th}}]\!]$ decreased in the above equation is a "witness" to σ's novelty (with respect to **A**).

2. This definition subsumes the $\mathbf{New}_{\mathbf{FI}}(\mathbf{Th}, \mathbf{A}, \sigma)$ condition. This follows from the fact that $[\![\{\}^{\mathrm{Th}}]\!](\mathbf{A})$ is equal to $\mathbf{I}^{\mathrm{Th}}[\mathbf{A}]$. (Note this $\{\}$ mapping is the "null mapping", whose domain is empty.)

3. Realize that if $\mathbf{A} \in \mathbf{Domain}[\varphi]$, then $[\![\varphi^{\mathbf{Th}}]\!](\mathbf{A})$ would contain a single member. As there are no non-empty proper subsets of such singleton sets, it is sufficient to use $\mathbf{Domain}[\varphi] \subseteq \mathbf{L} - \{\mathbf{A}\}$, rather than $\mathbf{Domain}[\varphi] \subseteq \mathbf{L}$.

4. Consider the set of "almost complete interpretations" $[\![\varphi^{\mathbf{Th}}]\!]$, whose domain includes every symbol of **L** except **A**; that is, $\mathbf{Domain}[\varphi] = \mathbf{L} - \{\mathbf{A}\}$. While it may appear that these partial interpretations are sufficient — that one of these would witness any new σ — the counterexample below shows that is not the case.

A tableau helps to visualize this definition. The left tableau in Figure 2 corresponds to the theory **Th3** ← {**A**⇔**B**} in the language **L** = {**A**,**B**,**C**}, and the one on the right to **Th4** ← **Th3** + "**A**⇔**C**". Each row is indexed by a mapping φ and each column by an assignment to **A**. A tableau position is tagged with a "1" if this assignment of $\mathbf{Domain}[\varphi] \cup \{\mathbf{A}\}$ is consistent — that is, if there is any full interpretation associated with this position — and a "0" otherwise.

Finding a witness to a sentence's novelty reduces to finding a row, r, in which a "1" is flipped to "0" but which does not vanish — that is, r must retain a "1" in some position. The fifth and sixth rows below (labeled with the "{<C,$\underline{F}$>}" and "{<C,$\underline{T}$>}" mappings) each satisfy this property, showing that $\mathbf{New}_{\mathbf{PI}}(\mathbf{Th3}, \text{"A"}, \text{"A}\Leftrightarrow\text{C"})$. Notice that none of the top four rows, which correspond to those "almost complete interpretations" has that property, demonstrating the point of Item 4 above. (These rows do form an adequate spanning set, though, as all the other rows can be derived by **OR**ing together appropriate sets of these.)

A: $\underline{F}$	A: $\underline{T}$					A: $\underline{F}$	A: $\underline{T}$
1	0	→	{ <B,$\underline{F}$>, <C,$\underline{F}$> }	←		1	0
1	0	→	{ <B,$\underline{F}$>, <C,$\underline{T}$> }	←		0	0
0	1	→	{ <B,$\underline{T}$>, <C,$\underline{F}$> }	←		0	0
0	1	→	{ <B,$\underline{T}$>, <C,$\underline{T}$> }	←		0	1
1	1	→	{ <C,$\underline{F}$> }	←		1	0
1	1	→	{ <C,$\underline{T}$> }	←		0	1
1	0	→	{ <B,$\underline{F}$> }	←		1	0
0	1	→	{ <B,$\underline{T}$> }	←		0	1
1	1	→	{ }	←		1	1

Figure 2: Partial interpretations for Th3 and Th4.

4 Uses

Even in its current unmechanized form, this definition can be used effectively as an analytic tool with which to understand many existing learning programs. Eventually, we hope to develop a "NewP" predicate or possibly a pair of operational (multivalue) functions: "Newσ", which generates $\mathbf{New}_{\mathbf{PI}}$ sentences from a given theory and symbol, and "NewSYM", which returns the symbols for which a given sentence is $\mathbf{New}_{\mathbf{PI}}$. This section lists several ways this definition (and its operationalizations) can be used.

- *Analytic Tool.* An adequate definition of newness would help us identify the sources (and recipients) of novelty within learning programs. For example, the teacher provides the ARCH program ([Winston 75]) with the new facts that enable it to learn. LEX's problem solver and critic are the sources of novelty for the rest of the system ([Mitchell 81]). AM ([Lenat 82]) has no clear source of novelty. This definition may also help us understand the distinction between compositional new terms — such as AM's definition of prime numbers — and other new terms, such as Bacon's use of intrinsic properties ([Langley 79]). Finally, it may lead to a definition of learning not based exclusively on performance.
- *Learning and Knowledge Acquisition.* An adequate (i.e., computable) definition of novelty might suggest ways of learning a topic more effectively. For example, it could focus the learner's efforts on those aspects of the domain where he has the greatest potential for acquiring something new. This information would help a knowledge-base builder decide which concepts need to be better understood, helping him to direct the dialogue. An analysis of a symbol's dependencies (defined above) might then be used to generate appropriate "probe" sentences to help understand this still vague concept.
- *Representation.* How should a given proposition be indexed? In general each concept should point to all the relevant facts that are *about* that concept. The most obvious approach, based strictly on lexical inclusion, is inadequate. For example, one would want to index "x + 1 = 0" by "x" and not by " + ", whereas "x + y = y + x" should be associated with " + " and not with "x".

  So how does one determine those concepts that a given fact is really about? We claim that "aboutness" is intimately tied to "newness" in the sense that σ is *about* a concept **c** whenever this σ expresses something *new* about **c** with respect to the appropriate diminished theory (which excludes g[s] and all of its consequences).
- *Linguistics.* The basic purpose of communication is for the speaker, S, to transmit a set of *new* facts, usually about some specific topic. To understand this process, we have to know what it means for a fact to be new to H and then how S (and H) can use this meta-fact when constructing (or understanding) the message.

5 Conclusion

While space does not permit an adequate discussion of the all the issues associated with this model of novelty, this paper would be incomplete if it did not address the following topics, and point the interested reader to the longer paper [Greiner 83].

- *Applicability.* The $\mathbf{New}_{PI}$ relation described above is applicable to any symbol in predicate calculus as well as propositional logic. In particular, the same formalism we saw work for constant symbols works adequately for relation symbols, albeit with an even larger tableau.
- *"Assertional Novelty".* The novelty we discussed above, $\mathbf{New}_{PI}$, is "definitional", in that its goal is to specify more precisely the referent of a given symbol. Another source of novelty comes from specifying some attribute of the concept; we label such facts "assertionally novel". (See [Woods 75].)

  These two categories are distinct: Imagine the symbol **RDG** had been totally determined, in the sense that the set I^{Th}("RDG") had but a single member. As such, there is nothing $\mathbf{New}_{PI}$ we can say about **RDG**. Despite this certainty, you still might not know what his hair color is. That is, **HairColor(RDG Brunette)** might be true in one interpretation, whereas others might hold that **HairColor(RDG Blond)**. Clearly **HairColor(RDG Blond)** is a $\mathbf{New}_{PI}$ fact about **HairColor**; however, most people would also want this it to be a new fact about **RDG** as well — that is, $\mathbf{New}_{Assert}$**(Th,"RDG","HairColor(RDG Blond)")**.
- *Intensional, not Extensional.* This paper has dealt exclusively with extensional phenomena, where novelty was determined with respect to the extensions of the symbols. Another approach is intensional — based on descriptions.
- *Deductively Closed.* Probably the most serious criticism of this work is its dependency on a complete deductive system and the requirement that each theory be deductively closed. New-sounding statements can also be used to focus the hearer's attention on some facts he already knew, rather than expose him to new facts. It should be possible to extend this formalism to handle such resource-limited deducibility. Then we could address topics like monotonic novelty and information obsolence.

Each of the issues mentioned above suggests a research task — that of plugging each limitation. The two issues we find most pressing are:

- Finding an equivalent but syntactical formulation of the semantical $\mathbf{New}_{PI}$ relation, in the same manner that $\mathbf{N}_{Syn}$ matched $\mathbf{N}_{Sem}$. We hope this will lead to one or more operational versions, of the types mentioned in the beginning of Section 4.
- Expanding this $\mathbf{New}_{PI}$ definition to work with deductive systems that are incomplete. (This reiterates the last issue shown above.)

Our basic thesis is that σ is a new fact about **A**, with respect to the theory **Th**, if, *under some set of circumstances, σ limits the number of interpretations of* **A**. $\mathbf{New}_{PI}$ achieves this by examining every partial interpretation, testing each to see if **A** loses a possible interpretation in that situation. This "partial interpretation" definition of context is clearly as general as possible. Furthermore, by counterexample, we have shown that this extreme generality is necessary.

Acknowledgments

We would especially like to thank Tom Dietterich for his significant contributions to this work. The following people also contributed: Professor Bruce Buchanan, Dr. Lew Creary, Jim Davidson, Dr. Johan deKleer, Dianne Kanerva, Dr. Jussi Ketonen, Jock Mackinlay, Dr. Robert Moore, Yoram Moses, and Ben Moszkowski. Many useful comments were provided by the reviewers. This basic research was funded by ARPA Contract #MDA903-80-C-0107.

Bibliography

[Buchanan 78] Buchanan, B. G., Mitchell, T. M., Smith, R. G. and Johnson, C. R. Jr.
Models of Learning Systems.
In *Encyclopedia of Computer Science and Technology*, . Dekker, 1978.

[Dietterich 81a] Dietterich, T. G. and Buchanan, B. G.
The Role of the Critic in Learning Systems.
Technical Report STAN-CS-81-891, Computer Science Department, Stanford University, December, 1981.

[Dietterich 81b] Dietterich, T. G. and Michalski, R. S.
Inductive Learning of Structural Descriptions.
Artificial Intelligence 16, 1981.

[Dietterich 82] Dietterich, T. G., London, R., Clarkson, K., and Dromey, G.
Learning and Inductive Inference.
In Cohen, P. and Feigenbaum, E.A. (editors), *The Handbook of Aritificial Intelligence*, . William Kaufman, Inc., Los Altos, CA, 1982.

[Enderton 72] Enderton, Herbert B.
A Mathematical Introduction to Logic.
Academic Press, Inc., New York, 1972.

[Gallager 78] Gallager, Robert G.
Information Theory and Reliable Communication.
John Wiley and Sons, Inc., New York, 1978.

[Greiner 83] Greiner, Russell and Genesereth, Michael R.
What's New? A Semantic Definition of Novelty.
HPP Working Paper HPP-83-26, Computer Science Department, Stanford University, February, 1983.

[Langley 79] Langley, Pat.
Rediscovering Physics With BACON.3.
In *6-IJCAI*, pages 505-507. Tokyo, August, 1979.

[Lenat 82] Lenat, Douglas B.
AM: Discovery in Mathematics as Heuristic Search.
In Davis, Randall and Lenat, Douglas B. (editors), *Knowledge-Based Systems in Artificial Intelligence*, . McGraw-Hill International Book Company, San Francisco, 1982.

[Michalski 83] Michalski, Ryszard S., Carbonell, Jaime G., and Mitchell, Tom M. (editors).
Machine Learning: An Artificial Intelligence Approach.
Tioga Publishing Company, Palo Alto, CA, 1983.

[Mitchell 81] Mitchell, Thomas M., Utgoff, Paul E., Nudel, Bernard and Banerji, Ranan.
Learning Problem-Solving Heuristics thruough Practice.
In *IJCAI-7*, pages 127-134. UBC, 1981.

[Moore 80] Moore, Robert C.
Reasoning about Knowledge and Action.
Technical Note 191, SRI Interational, October, 1980.

[Winston 75] Winston, P. H.
Learning Structural Descriptions from Examples.
McGraw-Hill Book Company, New York, 1975, chapter 5.

[Woods 75] Woods, W. A.
What's in a Link: Foundations for Semantic Networks.
In D. G. Bobrow & A. M. Collins (editors), *Representation and Understanding*. Academic Press, 1975.