Member-only story

In Search of a Universal Knowledge Representation

John Ball · Following
Published in Pat Inc · 8 min read · Feb 23, 2021

Knowledge representation is key to the future of natural language understanding because the right model enables all languages to share a common ‘repository of knowledge.’ But to this date, models are immature. By analogy, we haven’t seen the kind of breakthrough to better explain knowledge as Copernicus did in the field of astronomy. Fundamentally, models are misaligned with what we know in the cognitive sciences.

Today, I’ll look into the arbitrary nature of the current approach to knowledge representation as an enabler of artificial intelligence (AI) and consider an alternative optimized for human language representation. My justification for the alternative model? It is currently in use at Pat Inc. (Pat) for conversational AI use and it shows no sign of weakness. It enables generalization (i.e. common sense) and a reduction in pattern-matching effort.

This is an introductory article — more will follow soon to justify the theory.

Arbitrary Knowledge Representation?

Knowledge representations, such as entity-relationship diagrams and knowledge graphs, are not brain-like as studied in linguistics. They exclude human-like context and a general meaning-based representation. They are more like a database.

That's our starting point today. Few aspects of human languages are universal, and just because something seems evident in English doesn't lead to linguistic generalization. That's why I rely on Role and Reference Grammar[i] (RRG) for guidance since it was developed over time based on the analysis of diverse languages and their meaning.

In any case, human brains learn language automatically because language works the way brains function. They evolved together. The diversity of languages comes from the flexibility of brains in learning new patterns. There are a number of specific examples of brain function that will be explained in future articles.

An example of an arbitrary representation is seen in diagram 1, from Wikipedia's English Entity-Relationship page[ii] It contains a fusion between multiple concepts — a birthplace mixes location with the predicate (i.e. birthplace means "**where** someone was **born**"):

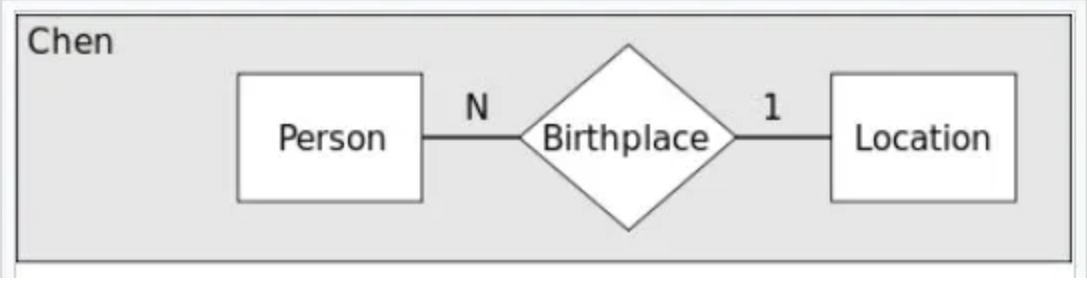


Diagram 1: An entity-relationship model from the 1970s, showing a flat relationship between an entity and a location, but 'location' is really more than just an entity.

Diagram 1 shows a person as an entity, the birthplace is a relationship between the person and a location, and a location is another entity. This approach inhibits scaling on slow machines, like human brains, because it requires additional effort to identify all the elements later. Locations are different from entities (referents): they are predicates.

> *Why a predicate? Take the question: "Where were you born" and convert it into a statement "You were born where". Now substitute for the word 'where' with your answer, say Iowa City. "You were born Iowa City." That's not English. Now add the predicate (preposition) 'in'. "You were born in Iowa City." Perfect. The predicate "in Iowa city" is a location — a relationship with the referent Iowa City. Language allows some words to be left out, like the answer to "Where were you born?" Answer: "Iowa City", but we know what it means,* ***in*** *the location Iowa City.*

What's arbitrary?

Selecting 'birthplace' as a semantic relationship, without regard to its meaning, is arbitrary. RRG models all of the world's languages with a single, layered model and is the opposite of arbitrary — it is consistent across languages. The layers look something like a dartboard when applied to semantics made up of elements that mainly answer questions like who, what, when, where, how and why.

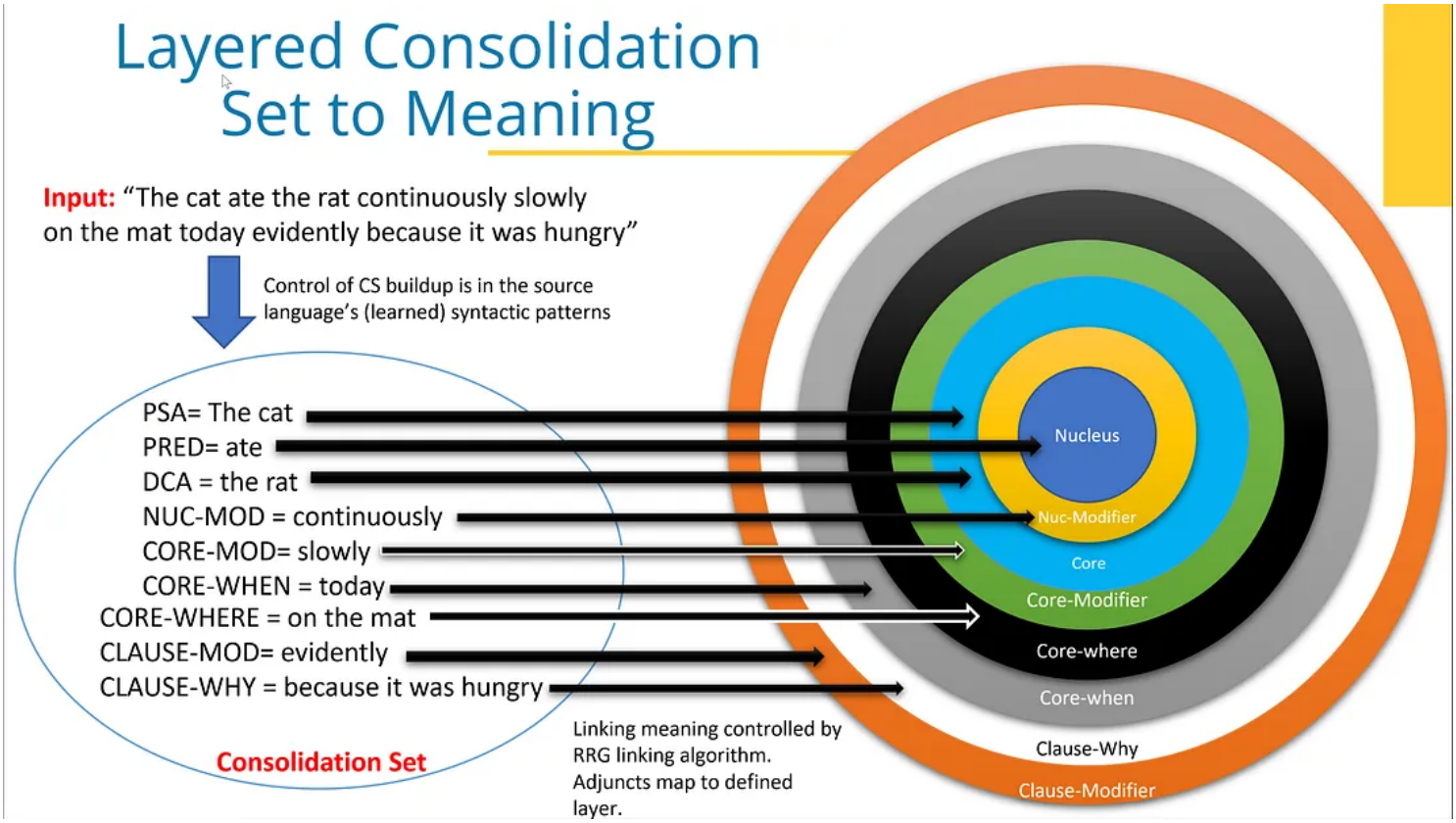


Diagram 2: Example shows a consolidation set (CS) of syntax (left) being mapped to the semantic set (SS) representation (right) — the RRG layered model.

In Diagram 2, the language shown is mapped to its layered semantic model. The two-step approach was pioneered by Pat in which the consolidation of the sentence takes place before its validation in semantics. Both are sets, in line with Patom theory's constraints that are based on observations of brain damage and other observations of function. There is flexibility in where many of the elements appear (in syntax) as their underlying meaning corresponds to a single layer. For example, 'today' is always a time (temporal) that answers the question "when did it happen?"

When we apply this layered model to the example in Diagram 1 above, the fusion becomes obvious. 'Birthplace' means the place of birth — a place that an event (a birth) happened. Instead of joining them in the semantic

diagram, keep them separated. In “John was born in Iowa City:” ‘born’ is in the nucleus, ‘John’ is in the core, and ‘in Iowa City’ wraps the combination as the location.

Just because English allows the combination of meanings in a single word doesn’t mean our semantic representation ought to copy it. After all, the brain may well have a layered model based on its localization of elements and some requirements (like anatomical brain connectivity) to keep their content separated.

Arbitrary Extension

The model can now be extended to make the point that the arbitrary combination of elements leads to an explosion in combinations. If you have a birthplace, you can also have a birthdate, a birth-month, a birth-year, a birth-city, and a birth-country. While a computer system happily allows the explosion in combinations because they are so fast, we should keep the model as simple as possible since brains seem optimized due to evolutionary pressures (societies typically value articulate people above those who don’t speak well).

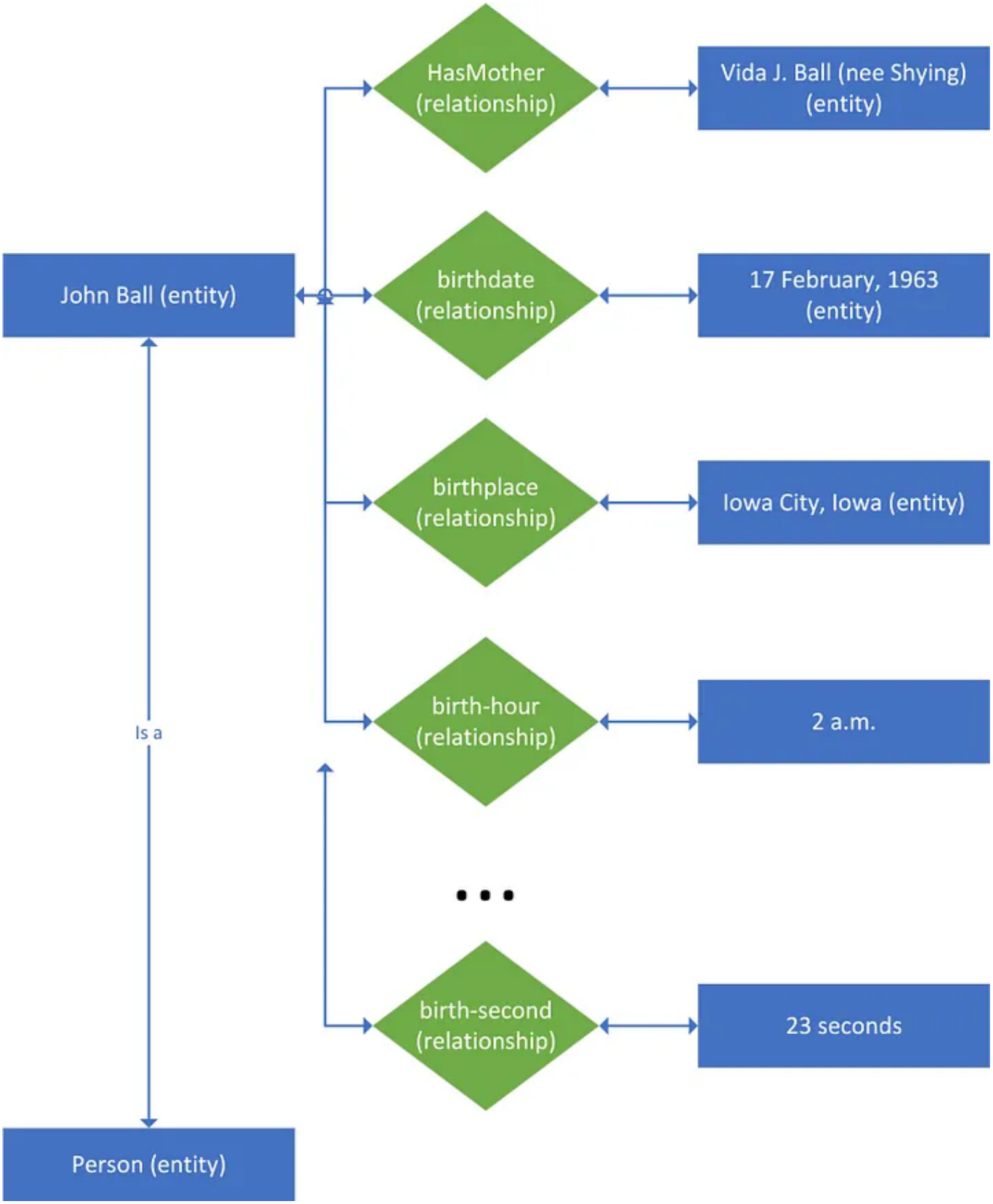


Diagram 3: Exploding information as often implemented today.

Not only does the arbitrary model allow endless extensions, but it is also difficult to compare questions as there are many answers in the same

category. When were you born? “At 2 a.m.” or “on Feb 17, 1963” or “in February”,… These answers are all valid, but they are disconnected: stored many times. Without context, they may be the wrong answer as well such as: “I was born at 12 noon. When were you born?” … answer: “on February 17.” (Nope. In context 2 a.m. is expected.)

The better approach is to align the semantic layers and store knowledge at whatever level of granularity is known. Brains don’t store in a DateTime format. 1963 may be all you know. Perhaps all you know is February 17. I mean, who knows what time of day your birth was made official! Software developers learned long ago that forcing developers to add code that never applies is a “code smell” (it’s bad) and similarly we shouldn’t add specific date information if unknown (e.g. Jan 1, 1963 at 12 noon instead of just 1963).

This is a key point — humans remember dates and times with as much specificity as is needed. Talking about Ancient Greeks can be done with a reference to 350 B.C., your birthday is normally expressed as the day and month, while talking about the next New Year is very specific as 00:00.000 on January 1, 2022.

Extending Further

The efficiency of the layered model is one thing. Another is its ability to generalize without effort. Inside the network, the meaningful elements have associations of their own, learned through language or other means. Much is written about referent associations (is-a and has-a, for example) and there are similar associations at the predicate level (decomposition and argument selectional restrictions). The representation stored may be awkward to convert into single sentences, but that doesn’t preclude storing in the most efficient manner.

Diagram 4 shows the same information in the form of the layered model:

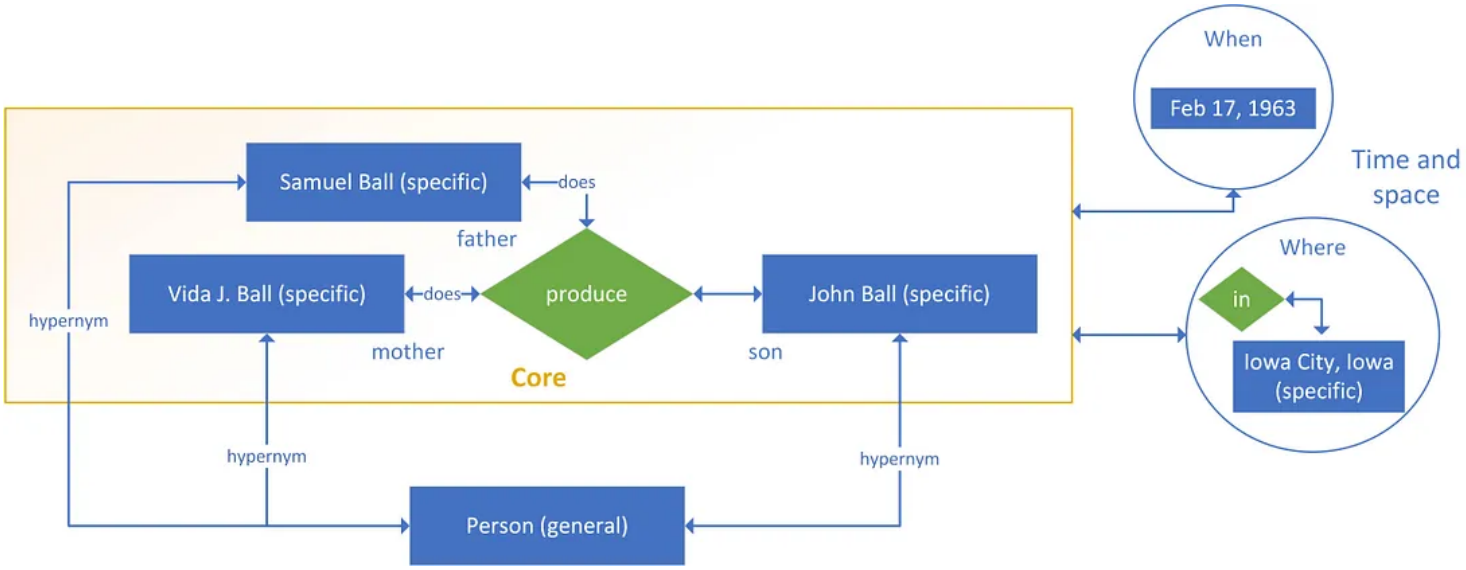


Diagram 4: Expanded information using a layered model. Each component can rely on its elements: the date includes month (Feb), day (17), and year (1963). The event is fixed in time and space in linguistics, a method that enables powerful generalizations and multi-linguality. This 2D diagram shows the wrapping as arrows to the collection comprising 'the core'.

Note that the *when* and *where* layers position the predicate and its arguments in time and space.

The same predicate meaning "produce" generates the words bear, born, fathered and bore in the right context in English. "Vida bore John", "John was born in Iowa," and "Sam was the father of John" are all paraphrases of the meaning shown in Diagram 4. To generate those sentences, the appropriate meaning elements would first be selected.

Many sentences can be created as a result of this one element of a larger knowledge graph:

- John was born in Iowa City.
- John was born in Iowa.

- Samuel Ball is John's father.
- Vida Ball is John's mother.
- John was born on February 17, 1963.
- Vida Ball gave birth to John Ball in Iowa City, Iowa, in 1963.
- A mother and a father give birth to a son.
- Some people give birth.

And so on.

Given a comparison time, the representation is sufficient with added tense, modality, and aspect to describe the situation accurately, as in:

- John will be born in Iowa City next week (future tense, if said on February 10, 1963)
- John should have been born in 1962 (modal, perfect aspect, if said today)

The event of the birth can also be extended indefinitely. The birth weight can be associated with the John Ball identified, which is already grounded in time. Many events can be connected to this predicate as a simple sequence. Representation is key to the dream of storing the world's knowledge in a single repository, independently to language — using meaning — for everyone to reference.

Repeating, once more, Simply

In the example above, a few pieces are considered in the relationship. Here's the comparison of the representations for "John is in Palo Alto, CA."

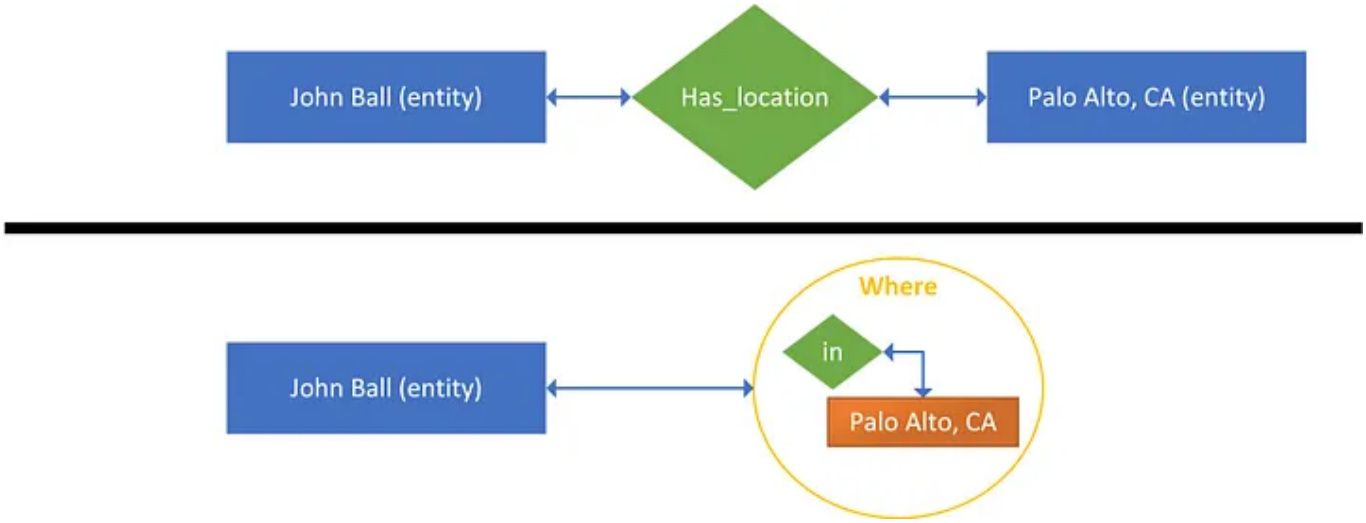


Diagram 5: The comparison between the flat entity-relationship model and a hierarchical model inspired by Patom (brain) theory. Note that 'where' logically wraps the entire core, which in this case is just a single proper referent. It's hard to draw.

In Diagram 5, the top shows an entity-relationship model and the bottom is the proposed one with the entity positioned in space by the predicate "in Palo Alto, CA." The Palo Alto referent has its own relationships, not shown here, but is more than just a language-independent keyword.

Using a hierarchy, instead of a flat model, enables a large number of sentences to make use of the same representation. And using the meaning of words (more on how this is done later) makes it possible to store and re-use knowledge between applications in any language.

In fact, instead of sharing *data* as is attempted today, the sharing of *knowledge* should be one of the immediate short-term goals that we set for AI since a language-independent resource can be re-used for any language without further processing steps.

Conclusion

The comparison between today's arbitrary representation and the one based on cognitive science has been discussed today. The development of a

language-independent repository of knowledge that is centrally stored promises many desirable new services.

In fact, meaning-based knowledge promises to be the **universal knowledge representation** of the future.

Modern linguistics is a tool that can improve the quality of our systems, especially AI systems. By considering the cognitive sciences, existing approaches in AI can be augmented with insights already studied in-depth.

Next time: what else a "super" knowledge graph can steal from linguistics since good artists copy and great artists steal. And a few legacy design decisions to overcome from our distant past that are still baked into our computer science.

[i] A complex but thorough book explaining RRG by Robert D. Van Valin, Jr.: https://www.amazon.com/Exploring-Syntax-Semantics-Interface-Robert-Valin/dp/052101056X

[ii] A reference about entity-relationship models: https://en.wikipedia.org/wiki/Entity%E2%80%93relationship_model

Nlu NLP Linguistics Cognitive Science Artificial Intelligence

More from the list: "NLP"

Curated by Himanshu Birla

Jon Gi... in Towards Data ...

Characteristics of Word Embeddings

· 11 min read · Sep 4, 2021

Jon Gi... in Towards Data ...

The Word2vec Hyperparameters

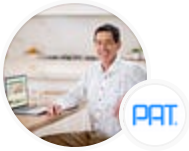
· 6 min read · Sep 3, 2021

Jon Gi... in

The Word2ve

· 15 min rea

View list



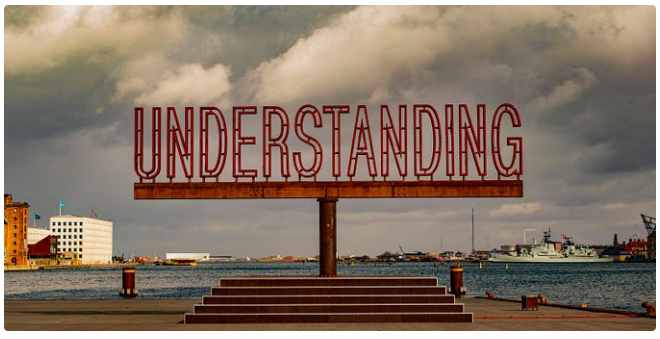
Written by John Ball

Following

1.7K Followers · Editor for Pat Inc

I'm a cognitive scientist working on NLU (Natural Language Understanding) systems based on RRG (Role and Reference Grammar). A mouthful, I know!

More from John Ball and Pat Inc



John Ball in Pat Inc

John Ball in Pat Inc

2023: The Year of Understanding Theory

Cybernetics was a name before AI was coined, and it is probably a better label for...

· 16 min read · Jan 5

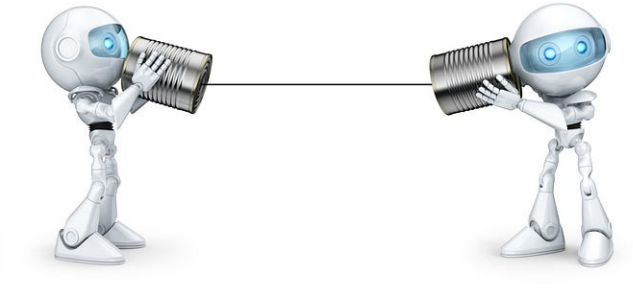
28 4

NLU and Consciousness

Natural Language Understanding (NLU) doesn't rely on consciousness, nor do most...

· 10 min read · Oct 28, 2022

99 7



John Ball in Pat Inc

Fixing Automatic Speech Recognition (ASR) with NLU

Why doesn't speech recognition operate at human-levels?

· 10 min read · Oct 22, 2022

94 2

John Ball

66 words handles 50% of language!?

At my company, PAT, we are working to replace today's systems with ones that...

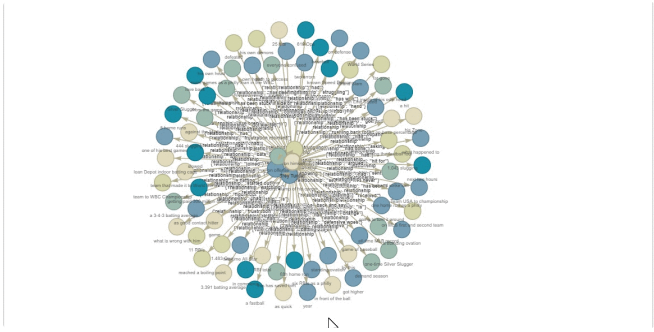
· 11 min read · Jul 6, 2022

107

See all from John Ball

See all from Pat Inc

Recommended from Medium



Wenqi Glantz in Better Programming

7 Query Strategies for Navigating Knowledge Graphs With...

Exploring NebulaGraph RAG Pipeline with the Philadelphia Phillies

✨ · 17 min read · 4 days ago

501 4

David Shapiro

A Pro's Guide to Finetuning LLMs

Large language models (LLMs) like GPT-3 and Llama have shown immense promise for...

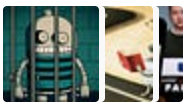
12 min read · Sep 23

283 6

Lists

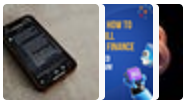
Natural Language Processing

669 stories · 283 saves



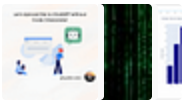
AI Regulation

6 stories · 138 saves



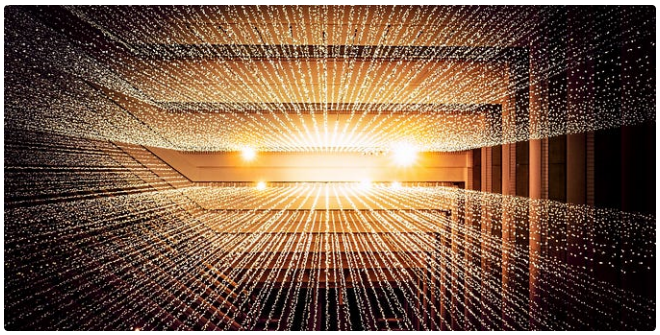
ChatGPT prompts

24 stories · 459 saves



ChatGPT

21 stories · 179 saves



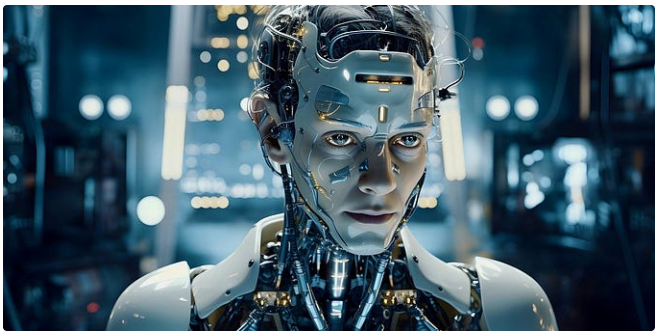
Beatriz Stollnitz in Towards Data Science

Add Your Own Data to an LLM Using Retrieval-Augmented...

Learn how to add your own proprietary data to a pre-trained LLM using a prompt-based...

21 min read · 4 days ago

189 1



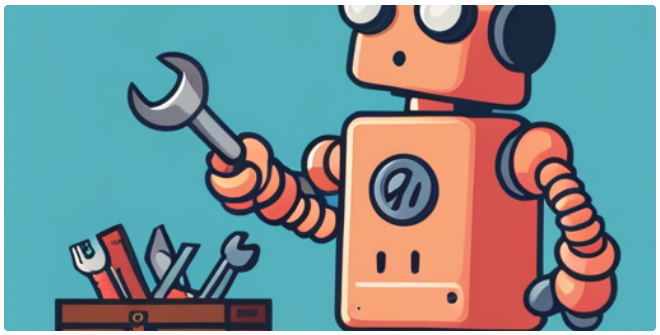
Nidhi Jain in Level Up Coding

Demystifying the Magic behind Large Language Models: The...

Step into the realm of AI architecture and uncover the designs that fuel the large...

· 10 min read · Sep 25

350



Chenhao Tan in Human-Centered AI

On AI Anthropomorphism

by Ben Shneiderman (University of Maryland, US) and Michael Muller (IBM Research, US)

20 min read · Jun 9

473 10

Dan Foster in In Fitness And In Health

In the Silence of the Hospital: Holding Hope for My Wife

The long wait in the shadows of the operating table

· 5 min read · 4 days ago

4.9K 112

See more recommendations