How to do Things with Metaphors:
Conceptual Science for Artificial Intelligence

Jongmin J. Baek
University of California, Berkeley
Department of Electrical Engineering and Computer Sciences
Department of Linguistics

December 16, 2015

Philosophical Assumptions

Look out the window to the highway. Businessmen in slick business suits wait patiently for the traffic light to change. Their horses lick the asphalt road and neigh. The smell of manure is overwhelming. The businessmen are clearly inept to ride the horses, and the horses do not enjoy the concrete road on their hooves, either. The traffic lights change, and it is at once pandemonium: one horse overtakes another, horses kick one another, several astounded businessmen fall off their horses, some more are trampled, and all the while, the horses are excreting in a continuous fashion that leaves the road absolutely drenched with manure.

This is what the current state of thinking about thinking looks like. The vast majority of the population still believe in Aristotle's 2500-year old theory, that of rigid, categorical cognition and Platonic essences. This theory is dysfunctional, arbitrary, and smelly, but most of the popluation still ride on it and hope to reach the sunset by the end of the day. What is worse, they have no conception that it might be wrong. They believe it is the most natural way of looking at the world. And amongst these delusions, the holy grail of human consciousness, metaphor, is unduly dismissed as specialized and poetic. Then, most of us do not know how we think; we do not know ourselves; we do not know each other; and attempts to replicate human consciousness on a machine, without knowing one bit about the object of the replication, is doomed to failure from the very start.

The scope of artificial intelligence is as wide and deep as consciousness itself. Unfortunately, it is dominated by applied statistics. Statistics give the illusion of the power to predict the future, but is in reality painfully limited. Google Translate, for instance, has collected literally all of the textual data in the Internet, but still cannot hold its head against a mediocre human translator. It is a race between the world's fastest horse and a F-15. What is the secret to the immense ability of human cognition that puts to shame the world's largest and most expensive supercomputers?

I had a frightening experience a few days ago. I was walking down campus late at night, when all of a sudden I heard an ununderstandable noise from behind. The only way I can characterize the impression is that this noise sounded like a dense, complex, ominous web, like sinister white noise. *It could be anything*, I thought, and froze. After a second, I realized the sound was just two people speaking to each other in a foreign language. I was dumbfounded. This was partly because I had just almost activated my urinary sympathetic nervous system, but what struck me as very strange was the fact that, when I heard something that did not sound like anything I had ever heard before, my default reaction was not curiosity but instead an overwhelming fear. More generally, I thought, perhaps it is the case that some experience *that could be anything*, an experience that does not map onto a past experience in any way, is the core characteristic of an experience that causes fear. This new experience does not map onto a past experience, but the brain attempts to map it onto one, tries to expand all possible connections (because *it could be anything*), fails, and the connection shrinks back into nothing. I hypothesize that we fear the unknown because the unknown cannot be mapped onto anything.

If you buy this argument, a computer cannot experience fear because it does not know how to make metaphor. In fact, a computer cannot experience any emotion because it does not know how to make metaphor. Conceptual science postulates that emotion emerges largely as a metaphor from the body, such as "he was flamingly mad", "she is a cold person", etc. But computers do not have a body, and this gives them nothing to make the baseline metaphors from. But what if we were to give a computer primitve embodied concepts, combined with a set of fuzzy, nondeterministic rules on making metaphorical connections from one concept to another, and feed it a lot of information?

I predict rampant disagreements here. To be sure, I am not asserting that it will take just some simple tricks to get a computer to think like a human being. I am not even saying that there is no fundamental hardware difference between the human brain and a computer. Decartes likened the brain to a water fountain, and later philosophers around the 18th and 19th centuries likened the brain to a steam engine, a factory, etc. The brain was continually likened not to what actually seemed like a brain, but to whatever was the most complex thing at the time. Maybe the latest buzz about computers' similarity to brains is just another data point corresponding to the linear trend of history, a perpetuated delusion that we are much simpler than we really are.

But I do not want to focus on the hardware, because I what I really want to say is that I believe that thought is a pattern, and it is not significant what the physical hardware upon where the pattern is manifested is, much like how *The Sun Also Rises* is significant in its words and not whether it is a hardcover book or an e-book. In fact, the words themselves are not even the significant part: it is the pattern of thought those words evoke that matters, and this is why a translated version *of The Sun Also Rises* into, say, Korean, is still considered a copy of *The Sun Also Rises*. Conceptual science predicts that thought begins with metaphor, the incessant chunking of primitive concepts to more abstract concepts, and an interplay from concept to concept through metaphor. Then is it not the structure of these metaphors that matter, not where the structure is instantiated in? What role do the stray neurotransmitters and chemicals play? Then, at least in principle, what stops these metaphors, these patterns of thought, from being instantiated in hardware other than the human brain?

The point is this. A statistical approach to artificial intelligence is destined for a dead end. Only by bringing in cognitive linguistics and conceptual science will artificial intelligence have a chance at breaking through the wall. In a vague, philosophical perspective, this makes perfect sense; true artificial intelligence is the singular attempt to replicate and transcend human intellligence, so how could it be achieved only through mathematics and science, when there is a whole another body of thought, namely literature and psychology and all of the humanities, that sits across the gulf? I see conceptual science as an attempt to bridge this gap, to merge the two great bodies of thought. Conceptual science declares that mathematics and literature are of the same breed, of the same root: that of primitive embodied metaphors.

In another such philosophical strain, conceptual science succeeds because it uses the scientific method born out of Western philosophy to confirm many aspects of what Eastern philosophy has postulated for a long time, thereby merging another two great bodies of thought. The crux of Eastern philosophy states that all is connected, that nobody is independent of others. In multiple East Asian languages, the "self" is linguistically realized as the collective self, the "us". For instance, it is customary to omit the "I" in Korean when one talks about herself; to say "I agree", one simply says "agree do." However, when one disagrees, one explicitly says the "I", such as “I disagree”. A frame analysis of the situation shows that the assumed frame is that of all memebers of the community acting as one and in agreement, and so the “I” of “I agree” is assumed and may be omitted. However, the “I” of “I disagree” is not present in the frame and must be explicitly stated. This analysis is a window to a collectivist society’s way of thinking: that all thinking is related, that we more or less think the same thing.

Conceptual science supports this line of thought in two ways. First, it establishes that all humans share the same primitive schemas whereby all thought is born. Second, it advances the argument that the only way a person Kiana can understand a person Yosef is through a metaphor from Kiana’s self to Yosef’s self. This metaphor must be instantiated via neural circuitry, but it is certainly not the case that Kiana thinks of Yosef as possessing exactly the same self as she does. Rather, the metaphor says Yosef is similar to Kiana but differing in some important ways, and by this extension, Kiana, having known Yosef, has had her own self extended. For example, if Yosef tells Kiana a heartwarming story of when he helped a starving homeless man by bringing him home and bathing him and feeding him, a lot goes on in Kiana’s mind. If Kiana thinks Yosef is phony, this story never gets attached to her “compassion” frame, and Kiana will probably add Yosef into her “phony” frame instead. If Kiana doesn’t think this and her heart is geuninely warmed by the story, this story is added to her “compassion” frame, and the story is already connected to the “Yosef” frame, which is connected to the frame of Kiana’s own self, so, in a way, Kiana’s self has added another experience of compassion. This modifies Kiana’s self in terms of Yosef’s self. Then Kiana, when confronted with some similar situation, can activate her Yosef frame to think what Yosef would do in that situation, and perhaps act accordingly. Over time it will be difficult to tell which part is Kiana and which part is Yosef. They have become literally inseparable in terms of brain circuitry. In the whole thought-as-pattern perspective, Kiana talking to Yosef can be viewed as an effort to translate a pattern in Kiana′s brain to a pattern in Yosef’s brain. And what is Kiana but a meta-congregation of those thought-patterns over the chronological dimesion? So Kiana is a pattern, and her thoughts at a particular moment are a three-dimensional cross-section of this four-dimensional pattern. So when Kiana transfers her pattern to Yosef, Yosef’s brain takes on what we hope would be a roughly similar pattern of Kiana’s. This pattern doesn’t just go away if the story is particularly poignant. In this way, Kiana and Yosef’s brains now share similar patterns, and if the only defining characteristic of a person is the pattern of her thought, then we can safely say that Kiana has integrated some part of Yosef into herself, and the two are inseparable. Much can be said about the moral implications of human interaction when viewed through the lens of conceptual science.

A Fluid Metaphor-Making Graph

Giving the computer primitive embodied concepts, while certainly nontrivial, would be the easy part. Really, the crux of the program lies on the formulation of a *set of fuzzy, nondeterministic rules.* How can one create such rules? And, after all, since all Turing machines must be deterministic, how is it even possible to design a nondeterministic algorithm?

We address the latter difficulty later. Let us focus on the creation of such rules first. For inspiration, let us take a look at the primitive embodied concepts: *journey, rotation, container, motion, contact, force, causation, result,* and so on and so forth. Conceptual science predicts the decomposability of each and every word into embodied concepts. The more abstract the concept, the deeper the hierarchy of metaphors will be, and the more concrete the concept, the lower the hierarchy will be. This means, at least in principle, it is possible to represent the lexicon of a certain human being as a graph structure, that is, a set of nodes and a set of edges that connect some or all of these nodes. These graphs can also manifest more complicated neural structures, such as a clique (a set of nodes where every node is connected to every other node) acting as gestalt nodes or frames, dynammic weight adjustment for disinhibitions and control nodes, etc. Edge weights can represent how far a concept is to another concept such that a low edge weight from concept “apple” to concept “red” means “red” is likely to be triggered whenever “apple” is triggered, and a high edge weight from concept “Christmas” to concept “electrical outlet” means one is not likely to trigger the other. Of course, it is impossible to know the structure of the graph nor the edge weights *a priori*, so it will have to be implmented by some sort of semi-supervised machine learning algorithm.

We will feed in primitive embodied concepts as the initial parameters for this algorithm. An army of linguists will have to label a set of sentences or pictures with these concepts, like the development process of FrameNet. Then we let the algorithm experience the world: we feed it pictures, Shakespeare, Buddhist scripts, movies, candy, coffee and sunlight. What I hope will happen then is that the algorithm will immediately be able to make mappings from the world to these concepts, then build chunks of these concepts for a slightly more abstract concept, then another round of chunking, then another, then another, *ad infinitum*. For example, the algorithm might see a man putting a piece of salmon in a tupperware. The schemas *container, journey, causation,* etc. are activated. The algorithm searches the web for what it has just seen, i.e. the tupperware, and finds that the semi-trasparent plastic object that fits the schema *container* is indeed named tupperware. A mapping is initiated from the image to the word.

We should be rather careful in this mapping. The sound of a word is very important, probably much more so than how the word looks. And the sound of a word has many connotations, influenced by the shape and motion of the human larynx when making the sound (section 1.9 of *Conceptual Science* describes this phenomenon). Moreover, in the human mind, the sound of a word is in itself a virulent metaphor: for example, the sound of “bicycle” evokes the sound of a similar word, say, “popsicle”, and once the mapping happens, unexpected overlaps and unexpected connections can take place: for example, *the bicycle has big popsicle wheels*, *When I was four my mom would let me lick a popsicle every time we went out to ride the bicycle together*, or even *A popsicle pops like a bicycle*. These sort of connections are not frequently made in normal human conversation, but using words for the sake of their sounds is a potent poetic device, and I have a hunch that there is something about the sounds of words that sways one’s cognition in a significant way. For example, using a lot of plosives and fricatives activates the neural circuitry responsible for manipulating the larynx to make such violent sounds, which in turn activate the neural circuitry responsible for violence. Lots of mellifluous liquids and nasals have the opposite effect. So the mapping from image to word should be from image to the sound of the word, rather than the text.

After this algorithm has been fed a large amount of data, it will have a huge lexicon with a structure similar to the human brain, with gestalt nodes, control nodes, x-nets, hierarchies of concepts such that more and more abstract concepts are tacked on more and more primitive concepts. Like a human, this graph should be continuously experiencing something, so that new concepts are continuously being added, new metaphors continuously develop, and edge weights continuously change. The “conscious” regions will be those where there is a particularly high amount of such activity. Maybe sleep can also be simulated if we cut off all stimulation and let the graph stimulate itself by gazing into itself, changing edge weights and creating metaphors based on itself. Once it has reached maturity, it will no longer be “fed” new data, but it will make its own choice on what data to look at, based on the configuration of its nodes and edges, based on what “interests” it.

What would interest a graph? As a human, I am interested in areas where a metaphor seem to be able to be made, but not quite, just out of the grasp. I think I am interested in these areas because once I make a metaphor from one to the other, I can understand something in terms of the other. This is why I am trying so hard in this paper to make metaphors from computer science to conceptual science. I believe this graph will be motivated in the same way, always seeking to strengthen edges that are very weak, but if built stronger, can connect two vast regions, thereby increasing its "understanding".

Soon this graph will gain a personality of sorts based on these bottlenecks. The graph will be "interested" in increasing its knowledge of concepts next to these bottlenecks, and will choose to consume data that activates these concepts.

Think of all the things such a graph would be able to do! It could have conversations: it could detect the embodied frames, such as agent and patient, of a statement or a question, activate the relevant frames, activate frames relevant to those frames, consider these activations, and put them together. For example, when faced with the question "Why are you staring at the sky?", this graph can detect the primtive embodied concepts "action", "journey", "goal", "direction", and so on, parse the question as asking for the "force" of this "action", and reply with something like "Because it has clouds.", the concept of "force" activating the word "because", the concept "journey" activating the concept "like", the word "sky" activating the word "clouds", and so on. And if the basic frame structures of all humans are similar, with their roots in primitive embodied concepts, what stops such a graph from being incarnated in multiple languages?

Machine Translation

The backbone of the state-of-the-art machine translation tehcniques is statistical learning methods, just like every other artificial intelligence algorithm. But as soon as we have a graph like this, we will be able to approach machine translation in a different direction. In this approach, translation starts with decomposing each word of the desired sentence or paragraph into its primitve concepts and retrieving its underlying primitive structures. Then we pour this molten essence into the mold of the desired language. First, we extract the frames and primitive metaphors that exist in a sentence. Since these are universal concepts, they must also exist in the graph of the target language. So we carry these over to that graph, and then it is a simple process of putting it all back together.

Poetry Generation

A graph as such could also write poetry. There are currently a number of poetry generation programs, none of them too serious, most of them relying on a simple probabilistic concept called Markov Chains. Particularly simple Markov Chains act under the assumption that the probability of a word occuring at some location depends solely on the word directly before it, and slightly more complicated Markov Chains act under the assumption that the probability of a word occuring at some location depends solely on the two, three, four, or *n* words directly before it. A Markov Chain algorithm can "learn" by being fed some text, and then it can generate some string of words according to the aforementioned assumptions. To give a feel of what this does, I have written a simple Markov Chain algorithm, fed it some of my own poetry, and asked it to spit some back out:

> They seep and leap and feel and all other
> mouths seemed occupied with drooling moist. Besides her bed, a bed
> full of lava. Why won't you melt inside of me? Your components are nice,
> and they looked at the sky and their faces grew
> red, and they sang in despair a lullaby rhythm. The
> rain, rolling by like guilty trains.

As you can see, Markov Chains are painfully limited and there is nothing interesting about them, nothing they can reveal about the poetic process, nothing about human cognition. But conceptual science can show us a better way.

What is a poem? What is style, beauty and grace in writing? Reasoning with the tools of conceptual science, I hypothesize that a stylistic piece of writing is one that knows exactly what frame its reader is in and can thus manipulate this frame in confident, unexpected ways, thus facilitating a maximum rate of communication and sparking sparkling, novel connections in the reader's mind. Consider zegumas; consider puns; consider the satisfying stylistic maneuvers exhibited by sentences that omit unnecessary repetitions; consider the assured posture of a sentence that gives the reader neither too much nor too little, is always just one step ahead of the reader, like a witty, teasing rabbit, jumping left, right, taking a deep breath, suddenly points at that little carrot at the back of your mind, gorges on it, and somehow, at the end of the sentence, makes the reader realize that the entire point of the sentence was the carrot, and that the reader will never look at that carrot the same way again. Boring sentences are boring because the idea the sentence tries to communicate is already manifest in a frame in the reader's mind. Interesting sentences are interesting because the idea the sentence tries to communicate is not manifest in a frame in the reader's mind. The previous sentence was probably boring precisely for that reason: the reader, after reading the previous-previous sentence, has in his frame the negation of that concept, a background for a figure, and describing the background after describing the figure is boring.

A useful heuristic in writing poetry, popularized by the poet Chad Davidson, is to omit "what comes for free." For example, one needn't say "the black night", for "black" comes for free with "night". The "black" is clutter. It exists, by default, in the frame of "night". Section 4.4 of *Conceptual Science* describes a similar effect: what is uttered is what contradicts the default frame.

Many middle school english teachers tell their students to stop writing the word "I". The students protest, not unreasonably, "but why not? I'm the one writing the essay, so why shouldn't I say "I"?" The usual answer is a demand for blanket submission to what your teacher says; the answer of conceptual science is that "the writer" already exists in the frame of "reading an essay", and so repeated invocation of the pre-existing element likely causes annoyance in the part of those teachers.

What does this all say about generating poetry? Poetry is the most beautiful form of writing: provocative, evocative, moving, and meaningful. A piece of poetry knows exactly what frame the reader is in from start to finish, and extends, manipulates, bends these frames to delight the reader. So a computer can generate a poem if it knows the exact frames of each word. Each word can be formulated as a vector, with words in the same frame being in similar vector distances, and a poem would be the successive invocation of concepts that are each just the right distance from each other; not too far, not too close, just enough to cause a burst in creative metaphor in the mind of the reader. This vector distance includes phonological distance as well as semantic distance, and in some cases the phonological distance will be privilieged to simulate the poet's process of using words for the sake of their sounds and finding unexpected connections.

Conceptual Science of Theoretical Computer Science

The algorithms used in this graph will all be search problems, which brings to mind the most profound problem in theoretical computer science: does P = NP? Here I want to set aside the graph for a moment and talk about how conceptual science could examine theoretical computer science, like how it examined mathematics in *Where Mathematics Comes From*. Attacking the fundamental questions of computability and complexity theory using tools of conceptual science is interesting because the computability theoretician would claim that computability can account for conceptual science and human consciousness, while the conceptual scientist would claim that conceptual science can account for computability theory (although I can imagine an alternative wonky scenario where the former explanation depends on the latter explanation and vice versa, a mutually recursive infinite loop of two of the most profound theories that attempt to explain consciousness). As the MIT theoretician Aaronson puts it, "If P = NP, then the world would be a profoundly different place than we usually assume it to be. There would be no special value in 'creative leaps,' no fundamental gap between solving a problem and recognizing the solution once it's found. Everyone who could appreciate a symphony would be Mozart; everyone who could follow a step-by-step argument would be Gauss; everyone who could recognize a good investment strategy would be Warren Buffett." So computability theorists believe that creating something is an algorithmic search problem. Writing a Mozart piece is hard because if the piece consists of a thousand notes and each note can vary in rhythm or tone in thirty ways, then there are a total of 30^{1000} pieces that can be made, an intractable number. But if there is only some way to efficiently search through this space and yield Mozart, that is, if P = NP, then there would be no value in creative leaps. For now, I want to take the perspective that complexity theory can be explained in terms of conceptual science, and claim that P is a metaphor for all that is conscious excluding the unconscious, while NP is a metaphor for all of consciousness, including the unconscious. If this seems out of nowhere, consider that a stroke of genius is rarely a conscious effort. Newton did not say one day, "I'm going to discover gravity today", and go on to do just that. Beethoven did not say one day, "I'm going to write the Fifth symphony", and go on to do just that. The stroke of brilliance occured in the unconscious, that is, the intractable space of NP-completeness. In this light, the P = NP question asks, is there a fundamental difference between the conscious and the unconscious? Is it possible, with no irony, to decide consciously to find a new mathematical theorem one day, and do just that?

P is the class of all problems for which there is an efficient way to find a solution. To consciously search for a memory, must it not be the case that there exists an efficient algorithm to carry out this search? For example, in the metaphor-making graph I described above, if the graph "wants" to think of a concept it must be able to get to it. For example, if it is "thinking" of the concept "dog" and wants to find all concepts connected to "dog" such that all these concepts are also connected to each other, this is a k-clique problem. As another example, if it wants to find a hundred concepts that are not connected to each other in any way, this is an independent set problem. Both problems are NP-complete, so we know of no efficient way to do this. If there is no efficient way to do this, perhaps the only way to find a concept lost in the midst of all other concepts, i.e. unconscious concepts, is through random free association, or through the mysterious power of metaphor. Maybe metaphor is a heuristic to search for solutions of NP-complete problems.

Conclusions

Over the span of ten digressing, disorganized and probably delusional pages, I attempted to paint a very crude stroke of what the merging of conceptual science and computer science might look like. Since I do not know much of either discipline, the painting should be taken with a grand grain of salt, and I would be flattered to see even a semi-serious consideration of any of its ideas. But I do have a baseless conviction that the mystery of consciousness cannot be cracked by just one perspective, such as the perspective of math, of literature, of psychology, of Western philosophy, of Eastern philosophy, of African philosophy, etc. I have a baseless belief that the wisdom of a culture or any serious discipline is the distilled essence of wisdom collected for millions of years, its basic sanity verified by evolution, and the fact that an idea has survived millions of years of ruthless evolution is a certificate of its profundity. As I write this essay, I notice a persistent theme that humans believe themselves to be less complex than they really are. In fact, this is a major topic of this paper, to reduce the brain to a graph in a computer! Maybe this is because we experience things one moment at a time, and any experience in just one moment does not seem too complex. Maybe we do not give ourselves enough credit, do not respect ourselves enough. Conceptual science is the only discipline I see that attempts to bring all of human wisdom under one umbrella without resorting to reductionist tools, and I believe this is why it is crucial.

Bibliography

Artificial Intelligence: A Modern Approach. Stuart Russell and Peter Norvig.

Algorithms. Dasgupta, Papadimitriou, and Vazirani.

Writing Poetry. Chad Davidson and Gregory Fraser.

Godel, Escher Bach: An Eternal Golden Braid. Douglas Hofstadter.

Where Mathematics Comes From. George Lakoff and Rafael Nunez.

Metaphors We Live By. George Lakoff and Mark Johnson.