Neural Networks 58 (2014) 131–147

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2014 Special Issue

Semantically-based priors and nuanced knowledge core for Big Data, Social AI, and language understanding

Daniel Olsher

Carnegie Mellon University, United States

ARTICLE INFO

Article history:
Available online 6 June 2014

Keywords:
Knowledge representation
Nuanced commonsense reasoning
Big Data
Machine learning
Social data
Natural language understanding
Data mining

ABSTRACT

Noise-resistant and nuanced, COGBASE makes 10 million pieces of commonsense data and a host of novel reasoning algorithms available via a family of semantically-driven prior probability distributions.

Machine learning, Big Data, natural language understanding/processing, and social AI can draw on COGBASE to determine lexical semantics, infer goals and interests, simulate emotion and affect, calculate document gists and topic models, and link commonsense knowledge to domain models and social, spatial, cultural, and psychological data.

COGBASE is especially ideal for social Big Data, which tends to involve highly implicit contexts, cognitive artifacts, difficult-to-parse texts, and deep domain knowledge dependencies.

1. Introduction

This paper introduces COGBASE, a statistics- and machine learning-friendly, noise-resistant, nuanced knowledge core for cross-domain commonsense, lexical, affective, and social reasoning.

Comprised of a formalism, a 10-million item, 2.7 million-concept knowledge base, and a large set of reasoning algorithms, COGBASE makes nuanced commonsense reasoning available to machine learning, Big Data, and social AI through the introduction of a *semantic prior*, allowing (potentially noisy) knowledge and models to accurately support concept-driven learning and understanding.

COGBASE's nuanced, primitive-based representation allows users to add new data, including conflicting data, without affecting existing algorithms.

Through the use of INTELNET Energy-Based Knowledge Representation (Olsher, 2012c, 2013; Olsher & Toh, 2013) and semantic primitives (discussed below), COGBASE is capable of representing a wide range of semantics, including nuanced commonsense world knowledge, narratives, emotion/affect, stereotypical situations and scripts, human goals and needs, culture (see for example Olsher & Toh, 2013), and the effects of context on reasoning.

COGBASE generates contextually-accurate expectations about the world, allowing systems to "fill in the blanks, reconstruct missing portions of a scenario, figure out what happened, and predict what might happen next" (Mueller, 2006).

In addition to providing theory, sample algorithms, and output for the capabilities laid out above, the present paper formalizes the key concepts of *nuance* and *semantic surface area* and demonstrates how reasoning and machine learning can benefit from knowledge systems maximizing these properties.

2. Benefits of semantics for Big Data and machine learning

Semantics represent an important frontier within machine learning (ML) and Big Data. While the overall need is clear (Breslin & Decker, 2007; Cambria, Rajagopal, Olsher, & Das, 2013; Cambria & White, 2014; Manovich, 2011; Murdoch & Detsky, 2013), questions remain regarding the problems semantics can solve and how they should be modeled and integrated with current approaches.

Without semantics, ML systems lose access to an important source of lexical information and implicit knowledge about the world. Semantics enable systems to relate lexical items that share no surface similarity (enhancing recall), to reject states of the world that are semantically inconsistent/'don't make sense', improving precision, and to make predictions about the world, enhancing performance overall. COGBASE is able to reason about the past and future, infer goals, decompose concepts, induce and test lexical item senses, gist documents, and much more.

Semantics facilitate identification of the real-world practical implications of lexical items, especially critical for social Big Data where inputs tend to assume significant shared context, much

E-mail address: dan@intmind.com.
URL: http://www.cogview.com.

http://dx.doi.org/10.1016/j.neunet.2014.05.022
0893-6080/

meaning is implied, and the presence or absence of a single lexical item in particular contexts may radically change overall conclusions.

COGBASE and INTELNET offer straightforward integration with natural language processing (NLP) and machine learning techniques, aiding deep reasoning. Semantics can assist greatly with sense disambiguation, opinion mining, reference resolution, and other key NLP tasks. Syntactic processing benefits as well; real-world social/Big Data texts are often ungrammatical or otherwise difficult to parse, and semantics (as demonstrated in the COGPARSE parser Olsher, 2012b) facilitate the identification of meaningful text spans and particular concepts of interest from which important information can be extracted.

Critically, *data domains interoperate under COGBASE*—data from one domain may be readily used in conjunction with information from another, and reasoning processes can straightforwardly consider data from multiple domains at once. As an example, a conceptual model could deliver INTELNET 'energy' (a form of information) to a spatial model, allow that model to perform reasoning, and then transfer the results back into the original conceptual realm. The structure of INTELNET makes cross-domain information transfers easy to visualize and to achieve in practice.

Big Data and social media content often involve opinion, culture, emotion, and other conceptually- and psychologically-mediated domains. COGBASE and INTELNET are especially well-optimized for data of this nature. As explained in Olsher (2013),

> 'Conceptually-mediated' refers to processes that unfold differently in practice depending on the specific concepts present in some knowledge base and on the specific ways in which those concepts are interconnected. An example would be moral perception, in which the specific concepts one has regarding virtue and vice and the implications attached to each of these all play a major role in determining how one will view a particular phenomenon. Similarly, 'psychologically-mediated' evokes phenomena which, in order to be successfully modeled, require reference to the functioning of specific psychological processes. A key example is word association; given the concepts 'sun', 'sand', 'ocean', 'water', and 'waves', a human would immediately reference and activate the concept 'beach', while a computer would *a priori* have no reason to do so. In cases like these, models with no capability to model semantic association would predict incorrect outcomes for incoming stimuli.

In summary, COGBASE and semantic priors allow ML systems to extract and make use of important new sources of information. Together, COGBASE and the associated COGVIEW formalism (Olsher, 2013) can model worldviews and commonsense knowledge, reasoning about both in an integrated fashion.

3. The COGBASE representation formalism

Given the myriad benefits, it is natural to ask why there has not been more widespread adoption of commonsense and semantic knowledge within Big Data, social data, ML, and natural language understanding (NLU).

One core issue has been that of *representation*. Traditionally, logic-based approaches have been employed in domains like those listed above. These approaches view knowledge as something expressible in the first order predicate calculus with a Tarskian semantics (McDermott, 1987), suggest that truth or falsity is central (and ultimately can be determined) and require the ability to decide whether certain statements ('logical sentences') are true or false. Deduction is the standard mode of reasoning.

Under logical methods, however, especially when considering commonsense, social, and other forms of non-propositional knowledge, important issues arise regarding *construal*, *nuance*, *implicitness*, *truth*, and *cross-domain model integration* (described below, see also Evans, Bergen, & Zinken, 2007 and Olsher, 2013 regarding the role of cognition in AI, NLP, and NLU). Questions have also arisen with respect to how commonsense knowledge may be profitably integrated with statistical and other methods. In this section we explore in-depth how COGBASE is able to address these topics.

In a general sense, the creation of knowledge involves the coalescing of otherwise undifferentiated stimuli into representable forms. COGBASE seeks to limit the influence that this extraction process exerts on the knowledge that is obtained and to *minimize the amount of assumed context that is unknowingly (and improperly) included*. This is important because the more knowledge is 'pre-construed' (see below) and pre-contextualized, the less flexibly it can support future reasoning operations.

COGBASE and INTELNET view knowledge as collections of experience and information that may be brought together, as needed and in a context-sensitive manner, to solve problems as they arise. Creative reasoning is greatly facilitated through the reuse of the same information in diverse ways across contexts.

COGBASE stores information at an intermediate level of abstraction (between symbols and connectionist networks). Knowledge is dynamically generated, based on the needs and restrictions of a particular context, through the combination of multiple 'bits' or 'atoms' of information. Atoms take the form of [*concept*, *semantic primitive*, *concept*] triples connected to one another within a graph database (presently Neo4J[1]).

Different than spreading activation, which traverses first-order predicate relations, INTELNET-based systems such as COGBASE involve the creation of new, highly contextualized concepts on-the-fly via the exchange of information *within* other concepts. Exposing the internal semantics of concepts makes it possible for AI systems to much more closely 'understand' what concepts represent.

The FACILITATE semantic primitive indicates that the presence of a particular item (i.e. *forks*) is likely to help facilitate some other goal (such as *eating*). Other primitives include SPATIAL ASSOCIATION, representing, for example, the notion that *students* are typically found around *schools*, TYPICAL, indicating that some set of semantics is prototypical for a particular concept, and STRENGTH, which modulates the degree to which one concept is expected to affect another.

COGBASE is designed to store many different types of data and information. Geolocation data, for example, is handled via a single unified map scheme, whereby various concepts are associated with particular points. In this way, proximity is made available as an input to reasoning.

Data arising from multiple domains may be represented within a single knowledge base and integrated quickly and easily because the core representation is very flexible and does not fundamentally change across domains. Each domain (spatial, affective, etc.) may require the definition of a small number of primitives unique to that domain, but all primitives interoperate through the same energy-based mechanisms.

COGBASE semantic primitives are designed to hide as little information as possible and are created at a level of abstraction intended to best facilitate real-world reasoning. When adding knowledge to the system, the theory always errs on the side of splitting meanings across multiple primitives, enhancing data availability. Information is coded with the intention of pre-cognizing (pre-interpreting) it as little as possible, (1) making it easier to reuse that knowledge in disparate contexts and (2)

[1] http://www.neotechnology.com/neo4j-graph-database/.

maximizing the ability of context to shape reasoning in appropriate ways.

Semantic primitives are intended to be as few in number and as semantically 'small' as possible, given that each additional primitive risks increasing opacity (a key quantity to be avoided). Database primitives are intuitive and easily understandable, making it possible to use human cognition to simplify tasks where appropriate by pointing the system towards specific knowledge subcomponents known to be useful for particular types of problems. Attention to those primitives most relevant to local problems and contexts enhances sensitivity.

Under COGBASE, the system is aware that particular atoms may not be dispositive of any particular question, may not hold in the present context, or may be completely incorrect. The idea, however, is that when a number of contextually-selected atoms are considered as a whole, they are capable of generating accurate knowledge and providing a powerful platform for intelligent reasoning about likely states of the world.

3.1. Nuance: the key to the kingdom

Of the concerns raised above, *nuance* underpins all the rest, facilitating the accurate modeling of social and other data, including that relying on complex contextualizations, deeply interconnected frames and concepts, and implicit reference to pre-existing shared knowledge.

Context and construal. In any knowledge representation, phenomena must be represented such that they may be viewed from diverse viewpoints cross-contextually. As an example, in a standard ontology TABLE would typically be represented as a type of FURNITURE, and reasoning would be based on this perspective (that is, a table can be bought at a furnishings store, it is something I would likely have in my home, and so on).

If it starts to rain, however, I must be able to reconstrue (change my viewpoint about) that TABLE, construing it instead in this context as a form of SHELTER. I can then reason using the latter viewpoint: if I am under the table I will not get wet, but if I leave I most likely will, other people may want to huddle underneath with me, and so on. *Any knowledge representation that only contains information about* TABLE *as* FURNITURE *will not be able to make the leap to the second perspective*, an issue termed *pre-construal*.

Another example of pre-construal is an entry the author once found in a knowledge base: *<country X> is a problem*. Clearly, such a statement can only be interpreted as narrowly limited to one particular context, intention, and perspective.

In order to make the 'messy' outside world fit into standard knowledge representations, traditional approaches often fit the world into a *standard* construal and encode that. This creates brittleness, however, because it is difficult to automatically adapt the resulting knowledge to new contexts. Such knowledge is also difficult to use as support for statistical methods, because it tends to only cover cases that have been strictly enumerated in advance, and statistical techniques are often brought to bear on novel (and noisy) data. Generally speaking, representation formalisms must allow access to enough 'raw' information to permit the generation of appropriate construals in specific contexts. It is always optimal to leave construal to runtime; COGBASE and INTELNET make this a computationally-tractable prospect.

Truth values. Traditional KR systems generally aim to define and discover truth values. In practice, however, and very often in the social world, truth is highly subject to context and probabilistic at best. It is often unclear what it actually means for a statement to be true or false. As an example, the question of whether CHOCOLATE is GOOD TO EAT hinges greatly on whom you ask and when. Dogs cannot eat chocolate, and, while many humans enjoy it, they are unlikely to answer this question in the affirmative after having eaten other sweet foods. There are generally no single answers to most social and many practical questions—these depend on the context in which a statement is interpreted, what has happened before, the attributes of the person making the decision, what a person considers to be delicious, what one might be allergic to, and so on.

Commonsense data can be impossible to codify in a logical manner and is often only partially correct or simply wrong (especially if the data comes from unverified sources). Moreover, real-world commonsense KBs can never be logically complete and commonsense reasoning is not monotonic in nature (cf. Todorova, 2006). Commonsense knowledge results from an incredibly wide range of interacting objects, upon all of which there are no a priori requirements in terms of coordination or similarity. It is impossible to maintain the consistency of one part of a database vis-à-vis any other when data is drawn from a wide range of domains and subcontexts that have *many concept interactions*, but *not many concept dependencies* that would push the overall system towards consistent definitions. This is especially true when data is not pre-construed and data from multiple contexts is mixed together; in such cases, contradictions are nearly assured (i.e. today is Tuesday only in the 'yesterday was Monday' partial context).

Opacity. Issues also arise with *opacity*—traditional KBs store data such that, beyond placing objects in relation to one another, all of the meaning of what is referred to is extrinsic to the database. The sentence 'The cat is on the mat' can be transformed into the statement *on(cat,mat)*, but none of the three symbols CAT, MAT, or ON contains any information about their deeper semantics, leading to the *frame problem*, or the inability to determine what remains constant when things change (and under what conditions) (McCarthy & Hayes, 1969; Shanahan, 1997). For example, if in the previous context we fill the air around the mat with catnip, the cat will likely not be on the mat for long, but this is no longer the case if we change the cat to a dog. In general, it has traditionally been difficult to predict exactly how reasoning should change when input changes, or to determine the general behavior of a concept under transformation without referring to some external source of information.

The deductive mode of reasoning. Deduction as a mode of reasoning requires strictly correct premises on which to base conclusions. Yet, often, such premises do not exist in the right form, they are wrong, or they are contextually inappropriate.

As McDermott puts it, "a deduction of the data is not sufficient [because] the requirement is too easy to meet. There will in general be millions of deductions leading to the observed conclusion, almost all of which are absurd as explanations. For example, one day … my clock radio was two minutes fast. … [T]here had been a power failure lasting two hours recently. One would therefore expect that the clock would be two hours slow, but I remembered that it had a battery backup clock. Hence, the proper explanation was that the battery-powered clock was inaccurate, and gained about a minute an hour" (McDermott, 1987).

In commonsense domains it is usually more important to reason towards that which can contribute to explanation, expecting noisy data that requires contextualization, than to deduce from given premises.

If we understand explanation as elucidating causes, context, and consequences, then it becomes clear that the COGBASE inference process is inherently well-suited to reason towards explanation, for the following reasons:

- it combines multiple pieces of information, all of which point to various aspects of causality, allowing the exact nature of that causality to become clearer as more and more pieces of information overlap,
- information is selected based on input context, and is thus more likely to point towards contextually-appropriate outcomes,

- once a concept is proposed, consequences can be readily determined and checked (see Possible Worlds inference in Section 10.1), and
- only those concepts that recur across multiple semantic atoms are selected, removing less-probable outputs and noise.

Multi-domain data. Lastly, it has traditionally been difficult to mix knowledge from different domains (spatial and conceptual, for example) because the representations for each domain are often quite different and there is no obvious way to determine how, say, spatial changes should affect conceptual data (a form of the frame problem).

Taken together, the above issues point to the need for an inherently nuanced knowledge representation, capable of working with noisy knowledge, performing contextualized deduction to the best inference, and avoiding preconstrual and frame problems, but that still remains tractable in practical cases. In the following section, these notions are made more concrete.

3.2. Formalizing nuance

As suggested earlier, of the concerns above, nuance is the most fundamental. This is because maximizing nuance in turn allows representations to avoid issues involving pre-construal, knowledge externalization, and symbol opacity. High nuance enables reasoning mechanisms that can handle noise, reason sensibly, and maximize the contribution of implicit knowledge. Nuance facilitates creativity by allowing systems to *reuse* knowledge differently across tasks – the very core of intelligence – and avoids the loss of domain-specific information during model building and domain knowledge translation.

There are four key indicators of nuance. First is the ability to dynamically construct a *contextually-appropriate version* of a concept, referred to here as $Concept_{in\text{-}context}$. If we conceive of concepts as 'fields of meaning' (as INTELNET does, see also Langacker, 1999), then both the generation of $Concept_{in\text{-}context}$ and the notion of context sensitivity translate into the ability to discover the most contextually-relevant information we have about particular concepts. As an example, consider the concept DOG. In a PET context, concept components such as MAN'S BEST FRIEND would best constitute $Concept_{in\text{-}context}$. In a CAMPING context, however, HUNT ANIMAL and CARNIVORE might be much more appropriate.

To formalize this notion, we may observe that, intuitively, there are two preconditions for the successful determination of $Concept_{in\text{-}context}$. First, the 'denser' the information generated by a particular representation scheme, the more content there is for an algorithm to select from during the contextualization process.

Second, there must be sufficient *surface area* (information exposed to easy introspection) within a graph to allow reasoning algorithms to extract maximal information. A representation must be built in such a way that context-relevant retrieval may access whatever information is available without that information being buried inside the structure of the formalism.

Concretely, we may define the *Surface Area for Contextual Influence*, or SACI, of some graph G as:
$SACI_G \propto \|concepts_G\| \cdot \|edges_G\| \cdot connectivity_G$.

Here, $\|concepts_G\|$ and $\|edges_G\|$ represent the number of concepts and edges in G and $connectivity_G$ is a measure of how densely connected the nodes within G are to one another.

We may usefully approximate $connectivity_G$ by the Beta Index, defined as $\beta_G = \|edges_G\|/\|concepts_G\|$.

If we then substitute this approximation into the SACI formula above, the $\|concepts_G\|$ terms cancel, and we are left with the result $SACI \propto \|edges_G\|^2$. This interesting result suggests that the number of concepts in a graph does not matter with respect to determining surface area; rather, it is the number of edges that counts, and exponentially so.

We may understand this as suggesting that, ideally, knowledge should be *highly distributed* across multiple primitives (i.e. multiple edges) instead of being concentrated within particular symbols.

Third, a nuanced representation must be able to support the generation of a maximal number of potential inferences (otherwise, the representation itself becomes a bottleneck). Maximal inferences occur when surface area is high, data is highly distributed, and primitives are sufficiently 'small' that a given concept generates many of them, making a maximal number of permutations possible. *It is important to note here that COGBASE does not perform any kind of search and is able to manage a very large space of permutations in a highly tractable manner* (see Section 6 for more).

More precisely, we may define the *Inference Generating Capacity* of a representation graph $\mathcal{G}$ as $IGC_{\mathcal{G}} \propto \frac{SACI_{\mathcal{G}}}{\sum_{i=1}^{|P|}\sigma(P_i)}$ where P is the set of edge primitives in use within $\mathcal{G}$, and $\sigma(p)$ is the semantic entropy of primitive p (defined next).

Semantic entropy, the amount of information implied by or contained within a particular primitive, can be understood by way of analogy to pixel size in images, with large semantic entropies corresponding to large pixels, and vice versa. As an example, the ConceptNet 4 relation *Desires* contains more semantic entropy than the COGBASE primitive *Facilitates*, because *Desires* implies a significant amount of contextually-grounded information about the needs and goals of a sentient actor, while *Facilitates* indicates just that a particular concept is often useful (in some unspecified way) towards the achievement of another goal/concept.

Substituting the definition of SACI into the formula above, we obtain:

$$IGC_{\mathcal{G}} \propto \frac{\|edges_{\mathcal{G}}\|^2}{\sum_{i=1}^{|P|}\sigma(P_i)}.$$

Thus, in order to obtain *maximal inference generating capacity* from a knowledge representation, we see that we must *maximize* the number of *edges* (primitives) across which information is encoded and *minimize* the *semantic entropy* of primitives.

We do not generally worry about primitives being too small, because there is no real penalty for using more of them in COGBASE, and smaller primitives facilitate more nuanced reasoning.

We may also define the overall *expressivity* of a segment of a representation as its average *IGC*. If the unit of analysis is the entire graph, then expressivity is equal to $IGC_{\mathcal{G}}$.

Finally, we may note that properly nuanced representation requires not just *small* primitives, but those of the 'right' semantic size to fit the data at hand. Continuing with the pixel analogy, if pixels are too large, 'blocky' images result that poorly represent the original source. Each overlarge pixel adds a significant amount of noisy information (termed *waste entropy* here) arising solely as an artifact of the representational system itself, biasing representation.

Mathematically, if we sum the squares of the waste entropy added by each primitive within some particular concept field, we obtain a useful measure of how well our representation is able to match the nuance present in the source domain. Of course, this measure implies that we have some way of accessing the original 'source'. Unlike an image, the only way to know whether we have found the optimal way of representing, say, the concept DOG, is to use our human judgment. If we consider a number of knowledge atoms that are intended to represent a particular concept, we may check that all of the important (to us) aspects are there and that, perhaps most importantly, we have not added anything extraneous by way of too-large primitives. We could run the category component decomposition algorithm (Section 10.4) in order to determine if the components returned there appear

sensical and whether or not anything significant has been added or omitted.

In summary, the key determinants of nuance ($\psi_{\mathcal{G}}$) may be combined as follows:

$$\psi_{\mathcal{G}} = \frac{IGC_{\mathcal{G}}}{\sum_{g \in \mathcal{G}} (\sigma(g_{represented}) - \sigma(g_{actual}))^2}, \quad \text{where :}$$

$\mathcal{G}$ is the graph for which ψ is calculated,
$g \in \mathcal{G}$ represents the individual concept-primitive tuples, or 'knowledge atoms', comprising $\mathcal{G}$,
$g_{represented}$ are the knowledge atoms as actually represented in the KB, and
g_{actual} are those atoms as they 'should' be according to a human oracle.

From the above, we may see that in order to maximize overall representation nuance ($\psi_{overall}$), we desire primitives with minimal semantic entropy, primitives that best fit the data, and graphs containing highly distributed information (with many edges).

The above precisely describes the design decisions underlying COGBASE. Primitives have been chosen in keeping with the wide range of semantics evidenced in the cognitive linguistics literature in order to provide the best fit for the widest number of scenarios (cf. Evans et al., 2007, Lakoff, 1990 and Langacker, 1999).

3.3. Inference

As outlined in Section 8, COGBASE depends on *energy- and data-guided probabilistic inference* as opposed to traditional methods such as modus ponens/tollens, offering a number of novel, important properties such as noise resistance (Section 4).

COGBASE allows knowledge from disparate portions of KBs to work together and allows for reasoning *within concepts*, permitting us to separate the various subparts of a concept and to reason independently about them (cf. Section 10.3.1). The idea is to enable 'computing at the level of the computer', whereby the system can mix and match semantic building blocks on-demand in order to meet dynamic task needs.

3.4. Intrinsic representation

COGBASE atoms offer a meaningfully *intrinsic* (cf. Waskan, 2003) form of representation in that some of the semantic structure of the world outside is mirrored within the database. This allows us to 'evolve' concepts and senses and to create new, contextualized concepts based on current needs.

Implicit knowledge is drawn from the interconnection patterns between concepts and the wider semantic atom interactions that these interconnections catalyze, as well as annotations on graph links, including semantic primitives, information about typicality, strength of expectations, and so on. The way in which any of these might become relevant during reasoning is determined dynamically based on knowledge and information needs at runtime, and indeed cannot be predicted until a particular contextualized traversal of the KB graph is undertaken.

Because COGBASE semantic atoms are easy to construct, and new knowledge implicitly benefits from old, the knowledge engineer need only insert relevant information about the most salient concept fields. It is not necessary to attempt to envision exactly which information will be needed or the ways in which that information might be used, as the system will determine this during runtime.

3.5. Semantic history and influence

COGBASE offers strong mechanisms for distributing semantic influence across reasoning processes and across time. As an example, during the processing of natural language texts, semantics are often expressed in the opening portions of dialogues which propagate to later portions. This includes argumentation strategies, introduced by the use of sarcasm or phrases like 'critics claim that', which tend to weaken the effect of following text. Also included are cases where certain concepts are made salient early on during processing and exert more influence than usual on future reasoning (for example, a topic sentence about *pets* might generate a context giving more importance to related concepts such as *dog*, *cat*, and so on).

Moral disapproval works the same way; when introduced in early stages of a dialogue, disapproval tends to spread to later concepts. Concepts discussed together are more likely to be disapproved/approved of together.

INTELNET energy provides a mechanism for representing semantic spread and modulating the semantics of knowledge encountered during processing. Such fine-grained semantics support opinion mining, perception modeling, and summarization tasks.

3.6. Frame problems

In COGBASE and INTELNET, frame problems are avoided in part by delaying full concept characterization until runtime, when sufficient context is available to change the course of reasoning. As an example, let us take the well-known 'gun in the oven' frame problem scenario. Normally, guns are capable of firing bullets. This is not true, however, if the gun has previously spent time in a hot oven, allowing it to deform. Traditionally, in order to determine this one would need to explicitly lay out all of the potential conditions and axioms under which a gun can and cannot be fired, a combinatorially difficult proposition.

Under COGBASE, however, the concept GUN (rendered in small caps to denote a concept space) would be not characterized until runtime, when it would become amenable to influence by contextual forces. If the system has knowledge that melting deforms objects, a GUN is a mechanical object, and that mechanical objects generally lose their function when melted, the system could infer that the main function of a gun may not be operative in this particular case. It could, for example, use the Category Component Decomposition algorithm (Section 10.4) to automatically discover that the concept SHOOT is the best concept handle for the prototypical result of the operation of a gun (in that this is the related action receiving the most INTELNET energy). It could then use a variant of the Concept Facet algorithm (Section 10.3.1) to remove data related to shooting from the GUN concept space. Reasoning could then proceed using this modified version of GUN, avoiding the need to explicitly specify axioms or conditions.

3.7. Further COGBASE design goals

COGBASE and INTELNET seek to make data *available*, meaning that it should be represented at a level of abstraction allowing maximal usefulness to reasoning (high surface area). All explicit and implicit deep semantics present in databases should be maximally exposed to the processes that run on top of them.

Data is *standardized*, such that an algorithm does not need to consider the source of information before drawing on it and algorithms need not be changed when new data is added. Performance should simply be expected to improve, as has been qualitatively borne out during development of the algorithms described below. Primitives allow fusion of data from different sources; after data becomes part of the system, it is irrelevant from which source it originally arose.

Data importation is *automated* as far as possible, so that once a translation has been decided between data source relations and semantic primitives, importation may proceed without further human intervention.

Lastly, while they are likely to be accessed through purpose-built software, the contents of the database are *comprehensible* via direct consultation. This is mainly achieved by selecting semantic primitives that are independently comprehensible, and by using a graph layout that is easy to visualize.

Holism. The philosopher Dreyfus (in Sun, 1994), suggests that traditional symbolic AI studies primarily deliberative rationality (i.e. analytical knowledge), grounded in Newell and Simon's physical symbol hypothesis. In his view, this has led to a state of affairs where AI has not yet fully accounted for intuition and situation-dependent reasoning, in which he suggests deliberative rationality must ultimately be rooted. For him, "[w]ithout [these factors], pure symbolic manipulation will not qualify as intelligence". He suggests that holistic, "holographic" similarity plays a large role in intuition and that, "[d]ue to [holographic similarity's] distributed nature, connectionist models may be better able to model intuition that symbolic AI fails to capture" (Dreyfus, 1992).

COGBASE is intended to provide a substrate wielding the power of connectionism, capable of calculating such "holographic" similarities and drawing upon them during reasoning. The multiple algorithms put forth in Section 10, especially those related to concept decomposition, characterization, causes, and consequences, represent a first step in this direction, providing an interlocking system of algorithms for calculating extended interactions between concepts.

Taken together, the above features offer strong support for advanced reasoning on social, commonsense, and natural language data.

4. Unique COGBASE properties: noise-resistance, gracefulness, openness to new data

The use of INTELNET and energy-based reasoning allows COGBASE to offer further unique properties.

Firstly, *COGBASE is highly noise-tolerant and noise-accepting*. Currently, the database contains a significant amount of incorrect and invalid entries arising from the original sources, yet it generates highly precise results. COGBASE's atomic design allows for techniques such as choosing the most commonly recurring semantics within particular contexts, traversing graphs based on task constraints, seeking similar semantics across multiple graph locations, selecting specific kinds of knowledge primitives (each of which embodies differing noise levels), and adjusting retrievals based on KB entropy (retrieving less data when entropy is high and vice versa), all of which, taken together, allow highly efficient noise reduction.

COGBASE allows *new data to be added without affecting old*. In traditional KBs, new facts often interact with pre-existing information in unpredictable ways, meaning that if new information is inconsistent, previously functioning queries may no longer continue to operate. COGBASE's use of a contextualized probabilistic inference system means that adding new information does not exert significant influence on pre-existing queries.

COGBASE reasoning demonstrates *graceful/gradual degradation in the face of noise*. In traditional KBs, a single incorrect fact is capable of generating arbitrary results. In many Big Data, complex modeling, and social media contexts, however, noise is ubiquitous and no particular set of assertions can be held to be correct.

COGBASE 'gracefulness' can be understood as gradual degradation such that *performance does not decline due to bad data if sufficiently accurate data is present elsewhere in the KB* until a majority of noise is present; even then, *inferences simply become only slightly, gradually less and less accurate*. In addition, bad data only affects inferences drawing on that specific information and is averaged out during data collection, so negative effects do not spread. The presence of inconsistencies is expected and accounted for during reasoning, and the system does not generate wildly inaccurate conclusions in cases where there may be relatively small errors. COGBASE algorithms are 'fail-safe' in the sense that, if they cannot answer a particular query, they will return nothing rather than provide erroneous information. It is therefore not necessary to sanity-check return values.

One way COGBASE achieves all this is to generally look for both evidence and corroboration of that evidence before making inferences. An example would be algorithms which consider information about what categories a concept is likely to participate in *together with* information about concepts participating in that concept as a category. In this way incoming category information provides evidence and outgoing information provides corroboration once the two are juxtaposed against one another.

5. The COGBASE knowledge core

Covering over 2.7 million concepts and 10 million pieces of information, COGBASE currently contains more than two gigabytes of data drawn from multiple sources, all translated into an INTELNET-based core representation.

The presence of a wide diversity of concepts in the knowledge base (KB) makes COGBASE effective for nearly any English-based task. Other languages can be added at a first level of approximation via the provision of cross-language links between lexical items. Even though lexicons may differ significantly between languages, the commonsense realities those languages describe do not differ nearly as much, making this an effective technique.

The KB can also be integrated with the COGPARSE Construction Grammar-based parser (Olsher, 2012b), which employs semantics during parsing to allow the extraction of information from grammatically-incorrect and meaning-dense documents.

As indicated earlier, COGBASE is organized according to a 'semantic atom' principle whereby observations about the world, including traditional logical relations (*Is A, Part Of*, etc.) are decomposed into smaller primitives which are then placed into a graph network. At runtime, atoms are bound together depending on task needs.

The knowledge base integrates directly with cultural, emotional, and social models along the lines of Olsher (2012a, 2013) and Olsher and Toh (2013), providing an immediate 'plug-and-play' knowledge layer.

While the current COGBASE KB is generated automatically from input sources, from a theoretical perspective COGBASE knowledge atoms are created by considering concepts pairwise and choosing the primitive that best describes how the first concept interacts with the other. As an example, when considering FORK and EAT, it is clear that FORK FACILITATEs EAT. This process is generally quite straightforward, making KB creation a low-effort proposition. Existing predicate calculus relations may be broken down into COGBASE primitives and then translated in an automated fashion.

In the KB, concept nodes act as 'handles' to the concept fields of individual concepts. Concept nodes appear only once for each concept-related lexical item per language, providing points of common contact across disparate data sources. Data for all senses of each lexical item is aggregated together, moving sense disambiguation tasks from the time of data input to reasoning, easing KB creation and facilitating accurate context-based sense disambiguation (see Section 10.3.1 for an example). If such disambiguation had been attempted at the time of data import, this would have limited the system to using default or most common senses, needlessly curtailing reasoning capabilities.

Wherever possible, the system makes maximal use of knowledge *implicitly present* in knowledge bases—that is, information that is not explicitly mentioned but which can be derived through the combination of multiple pieces of information or through the

creative reuse of existing information in new ways. This property acts as a 'knowledge multiplier', assisting in generating more intelligent behavior from lesser amounts of data and maximizing the potential number of inferences that can be made from the data practically available in any given context.

Data sources. Current sources include MIT ConceptNet 4 (Liu & Singh, 2004), DBPedia (Lehmann et al., 2014), CMU NELL (Carlson et al., 2010), SenticNet (Cambria, Olsher, & Rajagopal, 2014), and Princeton WordNet (Miller, 1995), including senses, gloss data with high-level descriptions of sense meanings, and the *See Also*, *Pertains To*, *Participle Of*, *Hyponym/Hypernym*, and derived relations.

COGBASE presently runs on top of the Neo4J graph database, with most algorithms written in Python and particularly performance-critical portions such as first-time database batch insertion and certain data retrievals coded in Java. The KB is accessible externally via a REST API (see Section 12).

6. Tractability and scalability: executes quickly, scales naturally to large knowledge bases

Even though they may potentially draw on gigabytes of data during reasoning, COGBASE/INTELNET algorithms can generally be straightforwardly optimized to run on standard commodity hardware with moderate RAM (10–16 GB).

One key reason for this is that, while in INTELNET all data is immanently available should it be required, in practice the reasoner only needs to consider a small part of the available space, and the representation itself makes it easy to determine what this space is without search. Specifically, contextualized energy flows guided by concept interconnections based on underlying commonsense semantics make it easy for the reasoner to determine what information to consider when. In essence, the reasoner does not need to 'think' in order to determine what data is relevant; the database has already implicitly performed this task in significant part by providing links between concepts that could affect one another. The need to catalog the potential diversity of interactions between concepts is in meaningful part handled via database structure.

Traditional deduction can be difficult to scale on large knowledge bases, because it seeks to determine everything that is *possible*. COGBASE and INTELNET, however, work to determine the most *likely explanatory* data, combining knowledge atoms within specific contexts in order to determine what is most likely to be true given knowledge of the world.

7. Putting COGBASE to work: semantic priors

A key contribution of the present work is the *Semantic Prior* (SP), which transforms COGBASE data into probability distributions immediately usable in machine learning and statistics.

A Semantic Prior implements the intuitive notion that, given the presence of particular concepts or items within a certain context, we can infer something likely about the past, present, or future state of the world in that context. An SP might deal with intentions and goals (for example, if I seek out a fork and knife, I probably intend to eat) or with the likely content of the world (if something explodes, I expect that in future some debris will result; if my oven is hot, someone must have previously turned it on, plugged it in, and so on).

The idea is that, given the world and the objects that appear within it, there is an inherent underlying 'commonsense prior' implicitly reflected in language and other AI-relevant domains. COGBASE allows us to begin to access this underlying distribution and to take it into account during processing.

COGBASE provides a family of SPs, each of which predicts within one particular *realm of prediction* (ROP). Each ROP in turn predicts answers to one fundamental problem of commonsense or semantic reasoning. As an example, the *User Goals* realm answers the following query: given that a user has shown interest in items *X* and *Y* (say, *fork* and *knife*), determine what goals the user likely intends to execute (*eat*, or *eat off of a plate*, for example).

More formally, a Semantic Prior (SP) is a function which, for some realm of prediction (ROP) $\mathcal{R}$, maps an input subset $\mathcal{C}_I$ of the overall set of COGBASE concepts $\mathcal{C}$ to a dynamically-generated probability distribution space (PSP) $\mathcal{P}_{\mathcal{RI}}$. That is, $\mathrm{SP}(\mathcal{R}, \mathcal{C}_I) \mapsto \mathcal{P}_{\mathcal{RI}}$.

$\mathcal{P}_{\mathcal{RI}}$ provides the probability that a certain concept will form part of an answer to the fundamental question posed by realm $\mathcal{R}$ under the context implicitly created by the input concepts $\mathcal{I}$. For example, if we let $\mathcal{I} = \{eat\}$ and $\mathcal{R}$ represent the 'state of future world' prediction realm, we might have $\mathcal{P}_{\mathcal{RI}}(\textit{lack of hunger}) = 0.8$. That is, if I eat now, it is fairly likely that afterwards I will no longer be hungry.

If we were to set $\mathcal{R}$ to 'Action → Emotion Prediction' and $\mathcal{I}$ to {*praise*}, we might then obtain $\mathcal{P}_{\mathcal{RI}}(\textit{happiness}) = 0.95$. Under 'User Goals', an $\mathcal{I}$ of {*fork*, *knife*} might generate $\mathcal{P}_{\mathcal{RI}}(\textit{eat}) = 0.98$.

The output of $\mathcal{P}_{\mathcal{RI}}$ is often used as input to further reasoning algorithms. Generally speaking, $\mathcal{P}_{\mathcal{RI}}$ will be highly sparse, in that most concepts in $\mathcal{C}$ will have (effectively) zero probability.

Theoretically, the set $\mathcal{C}$ is understood as consisting of all concepts present as lexical items in any natural language. In COGBASE, $\mathcal{C}$ is practically defined as the union of two sets: (1) concepts already present in COGBASE and (2) concepts provided within additional domain models. COGBASE already contains some limited technical knowledge, and domain models are generally only required in the case of highly technical domains (chemistry, physics, manufacturing, and so on). Current concept coverage is quite extensive and should be sufficient for most general problems. When required, domain models are easy to build, consisting of concept nodes straightforwardly connected to one another and to pre-existing concepts using standard primitives.

COGBASE concept node labels in $\mathcal{C}$ are not case-sensitive (practically speaking, all concepts are processed in lower case where this is sensible).

In the case of polysemous lexical items, data for all senses is connected to a single concept node (i.e. senses are not separated). Generally, the system reasons based on the most common sense (implicitly identified through frequency and ubiquity of common semantics). *Critically, however, the flexible design of COGBASE allows the data associated with particular senses as well as the semantic definitions of senses themselves to be automatically induced from the database*, and reasoning may be adjusted based on this (cf. Section 10.10). Specifically, atoms associated with the dominant sense may be suppressed if the system discovers that an uncommon sense is the most appropriate one in the current context.

Depending on the application, $\mathcal{C}_I$ might consist of concepts extracted from input documents, user queries, the output of another SP, or some other problem-specific set.

Each realm employs separate prediction algorithms based on underlying realm semantics and the kinds of COGBASE information that are relevant there. Depending on the specific primitives involved, one or more noise-reduction techniques may be employed.

As indicated above, COGBASE algorithms 'fail safely' in the sense that when information is returned, it can be trusted. Should insufficient data obtain within the database, or another error condition occur, no return value will be provided.

Table 1
COGBASE prediction realms.

– Possible Worlds – (*world that was, that will be*)
Given that concept $\mathcal{I}$ is present in the implicit context now, determine what other concepts are likely to obtain (or to have obtained) in ζ in the recent past or future.
– User Goals and Interests –
User Goal Inference Given that a user has shown interest in some set of items $\mathcal{I}$, determine what goals the user likely intends to execute.
User Additional Concept Interests (Search Augmentation) Given that a user is interested in concept $\mathcal{I}$, infer what other concepts/sub-aspects of $\mathcal{I}$ they are likely also interested in. Alternatively, given a search for $\mathcal{I}$, determine what other aspects of $\mathcal{I}$ are likely to be queried next. Adding these to the initial query should increase search quality.
– Psychological Model-Based –
Action⟶Emotion Prediction Given a particular action $\mathcal{I}_{\mathcal{ACT}}$, draw on commonsense and psychological data to generate likely emotion/strength outcomes if $\mathcal{I}_{\mathcal{ACT}}$ is undertaken with respect to a particular individual.
– Document-Level –
Gisting/Document-Representative Lexical Item Extraction Given a vector of lexical items $\mathcal{I}$, identify those most semantically representative of the entire document. This output can be viewed as a gist of the input document.
– Categories –
Category Component Decomposition Given a category $\mathcal{I}_{\mathcal{CAT}}$, determine what other concepts, collectively, generate the semantic prototype of that category.
Semantics-Driven Category Membership Determination Given a category $\mathcal{I}_{\mathcal{CAT}}$, find the probability that another concept $\mathcal{I}_{\mathcal{CONC}}$ would be considered a member of that category.
– Concept-Level –
Topological Concept Characterization Given a concept $\mathcal{I}_{\mathcal{CONC}}$, generate key semantic constituents by looking at nodes which have both inbound and outbound links to $\mathcal{I}_{\mathcal{CONC}}$.
Concept Intersection Given an $\mathcal{I}$ consisting of two concepts, generate concepts common to the semantics of both.
– Lexical Item-Based –
Automated Word Sense Induction/Membership Determination Induce the various senses of lexical items, returning the number of senses found and a distribution across the core semantic constituent concepts of each sense. This realm can also determine which senses best fit particular lexical items.
Document Topic Prediction Given the presence of lexical items $\mathcal{I}$ within a document, find the most likely topics of that document.

7.1. Current COGBASE realms of prediction

Table 1 briefly outlines the various realms for which COGBASE currently provides prediction algorithms. The goal here is to provide the 'lay of the land' for reference; extended discussion and sample outputs for each realm are provided in Section 10.

Each realm will find applicability to a wide range of machine learning and natural language processing tasks; in some cases, predictions will be useful for expanding the semantics of particular lexical items so that further regularities can be identified; in others, especially with respect to goal-related realms, the predictions themselves are sufficient to drive particular tasks.

In Table 1, ζ indicates a default context constructed anew for each COGBASE query, and $\mathcal{I}$ represents the input concept set.

7.2. Additional possibilities

COGPARSE integration. COGBASE data can be used to induce syntactic–semantic pairings from text which can then drive the COGPARSE parser (ideal for semantics and knowledge extraction from noisy text). COGPARSE employs knowledge during parsing, enabling the system to extract significant amounts of information for which syntax alone would not be sufficient (if correct syntax exists at all).

Under COGPARSE, each language requires a corpus of constructions (form-meaning pairings). Using COGBASE, these constructions can be induced from text in an unsupervised manner, termed *construction mining*. Under that algorithm, a set of unprocessed texts $\mathcal{I}$ is transformed into a set of sequences of semantic categories, which are then identified during parsing. The algorithm is quite effective; after only a small number of input texts common constructions such as '*the <object>*' can readily be identified.

Information extraction. Preliminary work has also been undertaken on an algorithm for determining the likelihood that a selected phrase in a document fits within a particular semantic category (such as 'Barack Obama' and 'President', or 'I went to France' and 'Travel').

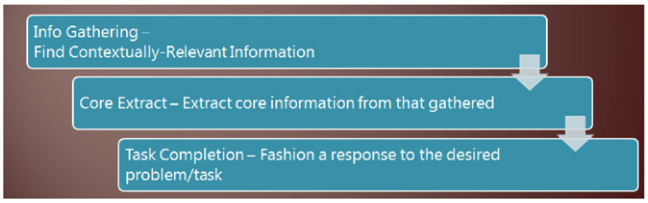
8. Reasoning with COGBASE

COGBASE reasoning processes are intended to quickly and efficiently discover, filter, connect, and synthesize contextually-relevant information from large, interconnected knowledge bases. COGBASE facilitates three main modes of reasoning: COMMONSENSE, COGVIEW, and HYBRID.

Sample algorithms are provided in Section 11 below.

8.1. COMMONSENSE reasoning mode

This mode is used most frequently with COGBASE. It consists of three phases, as follows:



The Information Gathering stage performs retrievals of particular concept and primitive data from COGBASE based on the contents of the input $\mathcal{I}$. Retrievals may be limited to edges in a certain direction/number of edge hops, and other conditions (such as shared inbound/outbound primitives) may be specified.

The next stage, Core Extract, executes a *core-generating function* (CGF) in order to build an initial set of useful information from the raw data gathered at the previous stage. A CGF might, for example, return the most commonly-appearing concepts in the data. Noise filtering and pattern detection typically also take place at this stage.

Finally, Task Completion transforms core information into a result acceptable for the task at hand, often by performing inference on the data contained in the core. See Section 11 for detailed examples.

8.2. COGVIEW and HYBRID reasoning modes

Interesting reasoning outcomes may also be achieved by combining COGBASE data with the COGVIEW worldview modeling formalism (the HYBRID mode), or by using COGVIEW reasoning with COGBASE augmentation. For more information on COGVIEW reasoning, see Olsher (2013).

One way in which these modes can work together for a conceptual input stimulus $\mathcal{S}$ is to simulate $\mathcal{S}$ through a COGVIEW network, collect intermediate and final concept energy levels, and then choose some subset of these concepts as input for COGBASE queries. This allows 'the best of both worlds'—integrated commonsense and social/psychological worldview models.

9. A note on bringing embodiment to computational systems

Embodiment – the notion that our experience as physical beings exerts significant influence on cognition and our understanding of the world – plays an important role in cognitive psychology, linguistics and robotics (Brooks, 1999; Evans et al., 2007; Pfeifer & Bongard, 2006), and has arguably affected the very development of mathematics itself (Lakoff & Núñez, 2000).

In practice, however, operationalizing this concept and integrating embodiment into computational systems can be difficult.

Much COGBASE data is inherently embodied in the sense that it encapsulates insights deriving directly from bodily experience (i.e. *hot* → *scald*, *burn*, *feel comfortable*, *intense*, *sweat*, *pain*, *ice* → *cool off*). It can also link various objects (*fork* and *knife*, for example) to the embodied goals they facilitate (such as using hands to fulfill the key goal of *eating*) via algorithms like those described in Section 10.2 below.

COGBASE is designed to maximize the ways in which a given piece of information can be used in diverse contexts, and can be adapted to support a large number of tasks, paving the way for it to act as an *embodiment enabler* for already-existing techniques.

10. Semantic prior output examples

In this section, examples are given of outputs for various COGBASE realms. In each example, for a specified concept/lexical item input vector $\mathcal{I}$, the 'output' set
$\mathcal{O} = \{c | c \in \mathcal{C}, \mathcal{P}_{\mathcal{RI}}(c) > 0\}$ is given.

Results are given as produced by the COGBASE system. In a very limited number of cases, some offensive or non-English result terms have been removed for publication, but outputs as given are accurate and have not been otherwise edited.

$\mathcal{I}$={*angry*}, Past/Future=Past
$\mathcal{O}$={*stub toe, read newspaper, jump up down, punish, punch, irritate, fight, person, watch television show, mad, fix computer, wait line, involve accident*}

Fig. 1. Past Inference: {angry}.

For the future (atelic) case:

$\mathcal{I}$={*sleep*}, Past/Future=Future, Telic/Atelic=Atelic
$\mathcal{O}$={*waste time, maintain health, rest, sleep, feel much lighter, death, re energize, rejuvinate body, refreshment, feel rest, wake up, rest dream, get better, rest body mind, rejuvenate body, restore mind, person feel better, snore, gain energy, lose job, awakeness, improvement health, energy, replenish neurotransmitter, refresh, bedsore, miss appointment, nightmare, not feel tire anymore, breathe problem surface, lazy, get proper amount rest, relaxation, no long tire, might dream, rest recharge, deep breathe, eat breakfast, escape world, lie bed close eye, wake up hungry, restore body, wake up fully rest, regeneration, feel better, take break work, bedtime, body feel comfortable, run out steam, not tire, pass hour dark, turn off light, drool, restore vitality, relax, rejuvination, baby, pass time, lie down feel sleepy, wake up morning rest, rejuvenation, slumber, fun, refresh mind, rejuvenate, recuperate, restful mind, become little tire, lie down, recover, refresh memory, delay, lucid dream, stay bed, feel energize, make tire person little tire, get rest, rejuvenatiobn, rest mind body, close eye, re-energize, maintain sanity, dont feel sleepy anymore, let body rest, satiate need sleep, become rest, dream, release energy*}

Fig. 2. Future-Atelic: {sleep}.

10.1. Possible worlds: Past and future

Given that a certain concept is salient now, this realm determines what some of the likely conditions are that could have given rise to this state of affairs. Similarly, given a concept in the present, it makes predictions about the future.

The Possible Worlds SP takes two arguments: Past/Future and Telic/Atelic (for Future queries only). Past/Future determines whether the algorithm is to look backwards or forwards in time. An Atelic query assumes that a particular action (EAT, for example) is still in progress and returns information relevant during and after the action is complete, while Telic queries are concerned only with what is likely to happen after action completion.
Fig. 1 shows the results of past inference for the concept ANGRY. For the Future (Atelic) case, see Fig. 2.

10.2. User goals and interests: Goal Inference

In this realm, $\mathcal{I}$ may consist of either a set of concepts or a single concept. In the case of a set of concepts ({*ham*, *bread*}, or {*fork*, *knife*}, for instance) the algorithm determines what goals the combined presence, use, or acquisition of the concepts in $\mathcal{I}$ is likely to support. $\mathcal{I}$ = {*ham*, *bread*} produces the probable concept set $\mathcal{O}$ = {*sandwich*}, and $\mathcal{I}$ = {*fork*, *knife*} generates $\mathcal{O}$ = {*eat food*, *eat food off plate*, *eat*}. With appropriate commonsense knowledge regarding terrorism, $\mathcal{I}$ = {*oil*, *fertilizer*} could generate $\mathcal{O}$ = {*bomb*}.

During processing, the system dynamically creates a 'mini-context' ζ from $\mathcal{I}$, and determines how the concepts in $\mathcal{I}$ interact with one another and with multiple potential goals under ζ. The semantically-structured nature of COGBASE removes the need for exhaustive search during this computation (cf. Section 6).

$\mathcal{I}$ may also take the form of a single concept representing an object, action, or state. For each case, the system generates appropriate probability distributions.

$\mathcal{I}$={*happy*} $\mathcal{O}$={*life go well, everyone, cat purr, score home run, good grade_fac_tn, child smile, person smile, find lose item, live life_fac_tn, enjoy day_fac_tn, everybody, pay, discover truth_fac_tn, get good grade, win baseball game, taste sweet, enjoy company friend, meet friend, gather energy tomorrow, money, party, celebrate_fac_tn, know healthy, happy, almost everyone, hear sing, surprise_fac_tn, mary sad mary, see idea become reality, read child, remember phone number, celebrate, buy present others, chat friend, love else, cash, download anachronox demo, love another, person, cheer, good lover, smile make person, enjoy day, mother, fun*}

$\mathcal{I}$={*fork, spoon*} $\mathcal{O}$={*eat food off plate, eat, eat meal*}

$\mathcal{I}$={*dog*} $\mathcal{O}$={*love, protect belongings, play, play frisbee, walk dog_fac_tn, bird hunt, run after ball, guard property, dog experiment, scare away bad guy, bark, find hidden, old lady walk, jog, walk dog, guide blind person, pat, companionship, guard home, help control livestock, companionship protection, breed, play frisbee_fac_tn, give comfort, comfort elderly, chase cat, track criminal, provide friendship, protect person, sniff out explosive, fetch newspaper, watch house, guard house, best friend, company, keep company, companion, protect livestock, guard junkyard, track animal, relaxation, sniff out drug, guard piece property, take walk, fetch stick, go walk, hunt, entertain person, person, pet, dog poop, fun, herd sheep*}

Fig. 3. Goal inference.

When $\mathcal{I}$ consists of a single concept, the algorithm interprets that concept as an object which has been acquired to help achieve some (unknown/unstated) set of goals and determines what those goals could be. The input set $\mathcal{I}$ = {*dog*}, for example, generates $\mathcal{O}$ = {*love*, *comfort elderly*, *protect belongings*, *play*, *guard property*}.

In the case where $\mathcal{I}$ contains a single action, the system assigns nonzero probability to goals which have that action as a component; the input $\mathcal{I}$ = {*kick*} returns $\mathcal{O}$ = {*swim*, *make mad*, *swimmer*, *fight*, *move ball*, *soccer*}.

Finally, in the case of world states (HAPPY, for example), the algorithm discovers goals that could have generated those states and/or that involve objects that can take on those states. In the latter case, the system may also return *facilitation nodes* (ending in *_fac_tn*) indicating specific actions that can be taken in order to generate those states. Examples are provided in Fig. 3.

10.3. User goals and interests: Additional Concept Interests, Search Augmentation

The prediction algorithm for this realm takes an $\mathcal{I}$ consisting of a concept in which the user is interested (perhaps the user has entered this as a search query) $\mathcal{I}_{INTEREST}$, an optional sub-concept facet selector concept (described below) $\mathcal{I}_{FACET}$, and parameters UseCategories, InCats, OutCats, ConfScores, and UseFacet.

During prediction, the system draws on KB knowledge to create a set $\mathcal{O}$ containing concepts which, given the user's interest in $\mathcal{I}_{INTEREST}$, the user is also likely to find important. As an example, given the search term $\mathcal{I}_{INTEREST}$ = *conference*, the user is likely to also be interested in terms like *workshop*, *speaker*, *keynote*, *venue*, *presenter*, and so on. This algorithm can be used in search augmentation; the set of search queries $\{(\mathcal{I}_{INTEREST}, C) | C \in \mathcal{O}\}$ should a priori be expected to collectively yield more relevant results than $\mathcal{I}_{INTEREST}$ alone.

When the parameter UseCategories is set to true, and either InCats or OutCats is also true, the algorithm expands the data search space using either the inbound (children → parent) or outbound (parent → child) semantic categories of which $\mathcal{I}_{INTEREST}$ is a member.

The parameter ConfScores determines whether or not the confidence values of the COGBASE data atoms from which $\mathcal{O}$ is derived are used to help determine final probability values.

In this realm each concept C in $\mathcal{O}$ is augmented with additional information about the number of times that C has appeared throughout the distributed data retrieved for $\mathcal{I}_{INTEREST}$, the aggregate confidence value of the information contributing to the probability value for C within $\mathcal{P}_{\mathcal{RI}}$, and an overall 'sort score' which is used to rank $C \in \mathcal{O}$ and generate final probability values.

Examples for $\mathcal{I}_{INTEREST}$ = *earthquake*, with UseCategories and OutCats set to true, and $\mathcal{I}_{INTEREST}$ = *terrorism*, with UseCategories, InCats and OutCats set to true, are given in Figs. 4 and 5.

This realm provides an excellent source of low-noise accuracy enhancement for general algorithms as well as data for concept semantic expansion.

10.3.1. Concept facets

When the parameter UseFacet is set to true, $\mathcal{I}_{FACET}$ specifies a *selector concept* used to intelligently narrow the results of data retrieval relative to $\mathcal{I}_{INTEREST}$. This narrowing can serve two use cases, *Sense Disambiguation* and *Concept Breaking*, detailed below.

Under both use cases, the system will automatically infer the semantic contribution of the selector term and determine the breadth of data that must be retrieved from the knowledge base.

Sense disambiguation. In this use case, a concept $\mathcal{I}_{INTEREST}$ with multiple senses is narrowed down to only one, specified by $\mathcal{I}_{FACET}$ (a single concept). An excellent example is BANK, which can refer either to an institution that manages money or to the side of a river. In this case, if $\mathcal{I}_{FACET}$ is money-related (*account*, *withdrawal*, etc.), that sense will be selected and $\mathcal{O}$ will be filtered accordingly.

Critically, *knowledge engineers need not specify which selectors correlate with which senses*; the system is able to use the totality of the knowledge base to automatically determine selector-data boundaries.

Concept breaking—facet selection. In this use case a single, complex concept with many facets is broken up and data related to one particular facet is selected for output in $\mathcal{O}$. In essence, $\mathcal{I}_{FACET}$ is treated as pointing to a semantic 'field' (range of interrelated concepts). As an example, the concept *China* refers to many things: a physical country located in Asia, a government, a people, various provinces and languages, and so on.

The selector term allows the user to choose which aspect of the larger concept they are interested in, and the system will automatically tailor knowledge to just that aspect.

As an example, with $\mathcal{I}_{INTEREST}$ set to *China*, an $\mathcal{I}_{FACET}$ of *government* generates the concepts {*govern*, *authority*, *money*, *organization*, *information*, *system*, *records*, *president*, *country*, *property*}.

With the same $\mathcal{I}_{INTEREST}$ and $\mathcal{I}_{FACET}$ set to *Asia*, we instead obtain {*continent, unite[d] state[s], nation, border, queen, america, origin, tropical country, continental area, popular destination, develop country, rapidly develop economy, earth, regional market, geography, property market, hong kong island*}.

From a natural language processing perspective, these capabilities provide programmatic methods for accessing the semantics and concepts associated with various lexical senses, allowing the construction of systems with much finer-grained semantic sensitivity.

$\mathcal{I}_{INTEREST} = earthquake$ (UseCats: TRUE, InCats FALSE, OutCats TRUE, ConfScores TRUE, UseFacet FALSE)
Augmented with data from automatically-generated outbound categories: natural disaster, physical phenomenon, earthquake

Concept	Count	Confidence	Sort Score
tremor	6 (3 orig)	2.55598	25.06611
shake	6 (3 orig)	2.52510	24.75222
event	2	4.36832	23.08221
mark robson	4 (2 orig)	2.72544	22.8561
seaquake	4 (2 orig)	2.66666	22.22222
disaster	7	2.43695	19.93874
emergency	2	3.28966	14.82188
hazard	2	3.11427	13.69869
catastrophic event	2	3.10550	13.64415
factor	2	3.08357	13.50846
tsunami	3	2.73037	13.45495
bring destruction house	3	2.73037	13.45495
natural phenomenon	2	3.04174	13.25221
shake ground	2	3.02979	13.17960
explosion	2	3.0	13.0
dynamite explosion	2	3.0	13.0
natural disaster	2	2.98788	12.92743
catastrophe	2	2.96957	12.81837
shock	3	2.59428	12.73031
emergency situation	2	2.95203	12.71451

Subsequent concepts include:

{disturbance, situation, fault, building collapse, ruin street pipeline house, person kill, bridge collapse, rift ground, many person die, fire, lot build collapse, many person hurt die, issue, risk, topic, calamity, threat, rumble, phenomenon, peril, traumatic event, incident, unexpected event, external event, circumstance, extreme event, problem, crisis, external force, unforeseen event, external factor, reason, unpredictable event, cataclysmic event, condition, force, unforeseen circumstance, unanticipate event, crisis situation, case, world event, natural evil, extraordinary event, extreme circumstance, natural process, danger, geological phenomenon, disruptive event, regional risk, tragedy, random event, emergency event, exclusion, environmental event, area, attack, disaster situation, scenario, item, emergency condition, evil, challenge, rare event, subject, environmental factor, large-scale incident, unusual event, sudden event, vibration, unforeseeable event, force majeure event, global event, tragic event, unplan event, feature, extreme case, outside factor, catastrophic emergency, national emergency, activity, localize event, concern, extreme situation, negative event, change, extreme, life-threaten event, story, emergency incident, warn sign, disruption, matter, field, effect, large-scale event, large incident, emergency circumstance, critical situation, one-time event, suffer, outside influence, cataclysm, unavoidable event, contingency, type, geological factor, safety issue, sudden shock, example, business interruption, civil emergency, unfortunate event, critical incident, difficult time, time, physical phenomenon}

Fig. 4. Sample results: 'earthquake'.

10.4. Category component decomposition

In keeping with the INTELNET/COGBASE view of concepts as having internal structure and being defined by combinations of and connections to other concepts, this realm uses KB data to identify a set of core concepts defining the field of a single concept of interest. The algorithm is especially useful in NLP (sense disambiguation, deep semantic processing), category matching, metaphor processing, and as part of most any algorithm concerned with concept and word meanings.

For this realm, $\mathcal{I}$ consists of a single concept, and $\mathcal{O}$ is a set of concepts which, taken together, can be considered to semantically recreate the $\mathcal{I}$ concept. An example is given in Fig. 6.

This algorithm also offers a low-entropy mode (used when data is especially sparse with respect to particular concepts in the database). Sample results are presented in Figs. 7 and 8. In Fig. 8, Concept Interests denotes the low-entropy version of

$\mathcal{I}_{INTEREST} = terrorism$ (UseCats: TRUE, InCats TRUE, OutCats TRUE, ConfScores TRUE, UseFacet FALSE)
Augmented with data from automatically-generated outbound categories: terrorism, bomb, intimidation

Concept	Count	Confidence	Sort Score
bomb	77	2.36322	92.58482
explosive	41	2.31045	56.33818
atom	10	2.24270	25.02974
war	9	2.52760	24.38881
blow up	3	4.03048	22.24476
state-sponsored terrorism	4 (2 orig)	2.66667	22.22222
domestic terrorism	4 (2 orig)	2.66667	22.22222
international terrorism	4 (2 orig)	2.66667	22.22222
chemical terrorism	4 (2 orig)	2.66667	22.22222
theoterrorism	4 (2 orig)	2.66667	22.22222
bioterrorism	4 (2 orig)	2.66667	22.22222
kill person	5	3.46548	22.00954
violence	7	2.30752	19.32463
issue	3	3.61484	19.06707
kill	5	2.56073	16.55733
intimidate	5	2.44444	15.97530
device	4	2.79599	15.81754
explosive device	4	2.76754	15.65928
threat	3	2.96595	14.79689
terror	3	2.95110	14.70899
destroy	3	2.93566	14.61808
element	4	2.55431	14.52448
weapon	3	2.91191	14.47920
topic	3	2.88196	14.30567
death	4	2.44595	13.98265
crime	3	2.73538	13.48230
fight war	3	2.73038	13.45496
event	3	2.68277	13.19721
destruction	3	2.66195	13.08598

Fig. 5. Sample results: 'terrorism'.

$\mathcal{I}$={*dog*}
$\mathcal{O}$={*dog, animal, pet, mammal, breed dog, cat, carnivore, species, often, bird, canine, popular pet, curious animal, home animal, noun, wild animal, plural cat, food, baby dog, small dog, household animal, live entity, domestic animal, common pet, domestic pet, feline, furry, man best friend, furry pet, pet animal, man, good pet, quadriped, domesticate animal, small animal, type dog, household pet, plant, young dog, predator, house pet, furry animal, person, golden retriever, beautiful animal, house animal, musician, rabbit, face other, hunt animal, ear*}

Fig. 6. Category decomposition.

$\mathcal{I}$=*hurricane*

natural element	emergency	sign
environmental condition	hazard	weather
extreme weather condition	consideration	physical factor
influence	natural force	natural phenomenon
idea	exception	circumstance
something	concern	effect
phenomenon	adverse weather condition	natural disaster
indicator	external factor	band
character storm	weather pattern	incident
weather condition	act	technology
image	type disaster	risk factor
risk	process	situation
	damage	

Fig. 7. Low-entropy category decomposition.

$\mathcal{I}$=*challenge*

Low-Entropy Category Decomposition	**Concept Interests (CI) (for reference)**	**CI Sort Score**
hard	run marathon	145
adjective	climb mountain	69
good	act play	66
problem	play chess	66
question	compete against	55
board	surf	55
game	take examination	54
truth	play hockey	51
situation	play game chess	48
difficult	puzzle	25
problem	challenge	14
easy	win	11
risk	fun	10
danger	competition	9
quiz	entertainment	8
trouble	dare	7
difficulty	appeal	6
chance	exercise	6
big	think	6
	demand	6
	(skill, practice next terms)	

Fig. 8. Low-entropy category decomposition.

the User Interests/Search Augmentation algorithm (included for reference).

10.5. Semantics-driven category membership determination

Accurate category matching is useful across a wide range of AI/NLP algorithms. In COGPARSE, as an example, the system must be able to determine whether various lexical items match specific categories present within linguistic constructions.

The Category Membership realm offers a semantics-based matching mechanism for determining the probability that a concept $\mathcal{I}$ would be considered as belonging to the semantic category $\mathcal{I}_{CAT}$.

The algorithm works for any concepts and categories for which a minimal amount of data is present in the knowledge base. As augmentation to the matching score provided as part of $\mathcal{O}$, specific information is provided on *why* items match, *how* they match, and *how well* they match, data highly valuable in metaphor processing and other applications.

Because category membership is determined semantically, matches can take place not only across traditional subcategories such as *chair* and *furniture*, which are most familiar to ontology-based modelers, but also via concepts such as *meat* and *tasty*, which draw directly on the deeper semantics of the concepts involved.

For example: $\mathcal{I}$ = {*meat*, *tasty*} generates an $\mathcal{O}$ containing the following two semantic comparison touchpoints: {*[food, 2.0], [animal, 1.73554]*} and a (very high) match score of 1.86777. These touchpoints, comprised of concepts and energy scores, indicate the shared core concepts which the categories and query concepts were found to have in common. Energy scores indicate the relative amount of semantic content shared by both concept and category with respect to each touchpoint. For match scores, anything greater than 1 represents a significant match.

The query also returns the following augmentation list illustrating the intermediate bases of comparison relied upon by the algorithm, together with energy values indicating the relative salience of each:
[food, 110], [animal, 100], [mammal, 50], [pork, 50], [beef, 40], [farm animal, 30], [bird, 30], [barn animal, 30], [lamb, 30], [goat, 30], [bone, 30], [chop, 30], [sheep, 30], [barnyard animal, 30], [ham, 30], [turkey, 30], [pig, 30]. Each concept listed is constitutive of the typical semantics of both the input category (*tasty*) as well as the specified lexical item (*meat*).

```
I am a relativist who would like to answer your question, but the way you
phrase the question makes it unanswerable.  The concepts of "right"
and "wrong" (or "correct/incorrect" or "true/false") belong to the
domain of epistemological rather than moral questions.  It makes no
sense to ask if a moral position is right or wrong, although it is
legitimate to ask if it is good (or better than another position).

Let me illustrate this point by looking at the psychological derivatives
of epistemology and ethics:  perception and motivation, respectively.
One can certainly ask if a percept is "right" (correct, true,
veridical) or "wrong" (incorrect, false, illusory).  But it makes little
sense to ask if a motive is true or false.  On the other hand, it is
strange to ask whether a percept is morally good or evil, but one can
certainly ask that question about motives.

Therefore, your suggested answers (a)-(c) simply can't be considered:
they assume you can judge the correctness of a moral judgment.

Now the problem with (d) is that it is double-barrelled:  I agree with
the first part (that the "rightness" of a moral position is a
meaningless question), for the reasons stated above.  But that is
irrelevant to the alleged implication (not an implication at all) that
one cannot feel peace is better than war.  I certainly can make
value judgments (bad, better, best) without asserting the "correctness"
of the position.

Sorry for the lengthy dismissal of (a)-(d).  My short (e) answer is
that when two individuals grotesquely disagree on a moral issue,
neither is right (correct) or wrong (incorrect).  They simply hold
different moral values (feelings).
...(signature)...
```

Fig. 9. Original posting.

10.6. Topological concept characterization

For a given concept $\mathcal{I}$, this realm generates an $\mathcal{O}$ containing concepts that are both the recipient of and originator of links to $\mathcal{I}$ within COGBASE (i.e. there are links in both directions). This realm provides a good approximation to the Category Component Decomposition (CCD) realm, is faster in some cases, and can sometimes provide results when CCD does not.

For example, given $\mathcal{I}$ = *fire*, $\mathcal{O}$ = {*cover, conflagration, blaze, blast, grate, burn, fiery, burning, ember, cinder, flame, light, fuel, ash, wood, smoke, heat, danger, combustion, spark, hot, something, heat source, harm, damage, burn hot, person, worker, sun, inferno, furnace, camp, fireplace, light match, burn wood, vehicle, power, house, water, department, earth, air, firing, rapid oxidation, huge fire*}.

For $\mathcal{I}$ = *perfume*, $\mathcal{O}$ = {*smell*, *scent*}.

10.7. Action ⟶ Emotion prediction

This realm predicts the emotions and perceptions that will arise when a particular action is undertaken with respect to another human being.

Drawing on the HYBRID reasoning mode (see Section 8.2), commonsense knowledge is used to determine how a psychological model will be affected by the input action, and the outcomes of that effect are then simulated by the system.

Energy values in Tables 2–4 are interpreted as relative strength values for each felt/perceived concept.

Concepts should be interpreted from the 'self' point of view—i.e. *Dominance* refers to dominance asserted against self by others.

Table 2
Energy distribution table for $\mathcal{I}$ = *Praise*.

Concept	Energy
Control	70
Good	70
Love	70
Power	70
Freedom	70
Comfort	70
Bad	−70
Live	70
Pleasure	70
Respect	70
Fear	−70
Care	70
Honor	70
Mean	−70

Table 3
Energy distribution table for $\mathcal{I}$ = *Insult*.

Concept	Energy
Love	−70
Power	−70
Face	−70
Trauma	70
Anger	70
Suffer	70
Mean	70

10.8. Concept intersection

Given two concepts $\mathcal{I}_1$ and $\mathcal{I}_2$, this algorithm determines other concepts which the two inputs have in common (that is, nodes that both $\mathcal{I}_1$ and $\mathcal{I}_2$ share links to).

article	better	right	motive	best
say	war	wrong	true	without
id	question	domain	false	position
like	need	question	hand	short
know	correct	make	ask	answer
two	answer	sense	good	two
person	something	ask	evil	individual
disagree	else	position	one	disagree
right	short	right	ask	issue
one	nice	wrong	question	neither
wrong	hope	ask	answer	right
sometimes	tell	good	consider	correct
rarely	assumption	better	judge	wrong
pretty	value	position	problem	incorrect
good	real	point	agree	hold
idea	statement	look	first	value
one	like	perception	part	john
wrong	assumption	motivation	position	department
never	value	one	question	state
information	part	ask	reason	state
make	objective	right	state	behavior
best	reality	correct	one	sort
really	like	true	feel	get
must	answer	wrong	peace	drink
make	question	incorrect	better	pick
decision	way	false	war	write
idea	phrase	make	make	part
right	question	little	value	life
judgement	make	sense	bad	
peace	concept	ask	better	

Fig. 10. Gisting results.

Table 4
Energy distribution table for *Respect*.

Concept	Energy	Concept	Energy
Dominance	−4071.5	Power	2770
Offended	−1670	Equality	4071.5
Conflict	−1120	Life	4071.5
Shame	−1110	Freedom	4071.5
Aggression	−721.5	Respect	4071.5
Mean	−560	Face	6301.5
Anger	−560	Respect for country	6301.5
Control	550	Country	6301.5
Care	550	Safety	6851.5
Authority	560	Values	9803.0
Hospitality	560	Happiness	13 153.0
Polite	560	Core emotions	13 693.0
Love	560	Core needs	14 273.0
Joy	1110	Sociality	17 204.5
Trust	1660	Social inclusion	18 314.5
Honor	1670		

As an example, for $\mathcal{I} = \{acid, base\}$, we obtain $\mathcal{O} = \{$*theory of dissociation, aqueous liquid, reaction parameter, bile salt, chemical liquid, inorganic chemical, electrolyte, ammonia, conductive material, reactive chemical, environment, program, fuel, ingredient, mixture, combination, material, chemical concept, deamination, reagent, compound, desirable quality, chemical substance, term, function, traditional general chemistry topic, form, brand, catalyst, constituent, raw material, list material, key word, oxidize agent, stabilizer, inorganic catalyst, volatile compound, agent, ionic compound, topic, volatile organic compound, harsh condition, feature, chemical, parameter, product, object, ph modifier, optional component, chemical compound, water treatment chemical, ionizable compound, class, alcohol, ionic species, chemical additive, liquid, metal, element*$\}$.

10.9. Utility function: concept semantic specificity

This utility function, calculated based on the ratio of inbound to outbound category links, determines how *specific* a particular concept is.

For instance, *place* (semspec 0.00314) is less specific than *United States* (semspec 11.0).

10.10. Automated word sense induction/membership determination

This realm covers word senses; COGBASE knowledge allows both the *automated discovery and induction of word senses* as well as *semantic sense membership checking*.

For the concept *mouse*, for example, the system is able to discover that there is one sense involving a computer product and another involving a living, moving creature.

The system is also able to check which of a number of senses a particular word usage is associated with (currently in beta testing).

{*small, mammal, animal meat, physical measure, hunter, chew, carnivore, companion animal, cat chase, nonhuman animal, common fear, routine surgery, sign, broad subject, high risk occupation, street, design, guard animal, common household pet, non-native predator species, coyote, old lady walk, untrain cage animal, better sense smell person, sausage, product category, grown puppy, short-lived animal, resident pet, move part, small horse, unwant animal, live entity, paw instead hand, go out better let, unwant visitor, suspect, companion, farm animal, new, faithful pet, non-human animal, simple piece, bird, hairy, animal model, interest group, leg, eutherian*}

Fig. 11. Sample naive outbound raw features: {dog}.

{*relate wolf, spot movement, scratch, bear, zoey, bichon frise, chow chow, robin hood, mother young, psycho, old english sheepdog, smell scent, muzzle, participate game, great pet, obedience, irish setter, shar pei, smell drug, circle house, hear many noise human, fetch object throw, tail, blaze, learn trick, dog show, carrry mouth, guard premise, breed several puppy, airedale terrier, use pen, run fetch frisbee, guard house, gina, cocker spaniel, mother puppy, coursing, listen sound, jack, ration_fac_tn, akita, pariah dog, train catch frisbee, malamute, die, marry, gsd, wire-haired, smell well*}

Fig. 12. Sample naive inbound raw features: {dog}.

10.11. Gisting/document-representative lexical item extraction

Given a document, this realm extracts those lexical items most likely to be semantically representative of the document as a whole. It discovers which semantics recur throughout and then selects only lexical items including those semantics, thus using the document itself as a base for filtering. This provides accurate *semantic gists* of document contents, with the frequency of individual lexical items within the gist indicating the importance of those words to overall document semantics.

In this realm, $\mathcal{I}$ is defined as a vector containing the lexical items contained within a single input document. If a given lexical item appears multiple times within a document, it should also appear the same number of times in $\mathcal{I}$ (that is, multiplicity matters).

As an example, with $\mathcal{I}$ set to the newsgroup posting[2] in Fig. 9, the $\mathcal{O}$ presented in Fig. 10 is generated as output.

Fig. 10 can be further compressed by counting the frequency of each lexical item present therein, as follows:

{*moral: 6, ask: 6, question: 5, right: 4, make: 4, wrong: 4, position: 4, certainly: 3, better: 3, state: 3, one: 3, answer: 3, good: 2, implication: 2, degeneracy: 2, correct: 2*, ...}.

Given that the input article is about moral relativism and ethics, both gists are highly accurate.

10.12. Document topic prediction

For an input vector of document-derived lexical items $\mathcal{I}$, this realm determines the concepts most likely to describe the topics present in $\mathcal{I}$.

One method of achieving this involves extracting semantic features from each lexical item in $\mathcal{I}$ and then applying clustering methods, such as Group-Average Agglomerative Clustering (GAAC), to the result (presented in Rajagopal, Olsher, Cambria, & Kwok, 2013).

10.13. Polarity augmentation

While COGBASE offers reasoning-based methods for opinion mining (cf. Olsher, 2012c), COGBASE data may be used to augment concept polarities, extending concept coverage and enhancing contextual accuracy.

10.14. Raw semantic feature generation

COGBASE data can facilitate the generation of raw semantic features from concepts and lexical items.

One naive algorithm for generating such features is simply to collect the COGBASE graph neighbors for each input concept. Under this method, however, noise is reproduced unchanged, accuracy enhancements are not performed, and primitives are not taken into account (thus generating mixed output semantics).

Outbound graph edges generate features through which input concepts define themselves via reference to other concepts and vice versa.

Examples are presented in Figs. 11 and 12.

11. Sample algorithms

In this section we consider two sample COGBASE algorithms under the COMMONSENSE reasoning mode (see Olsher, 2013 for information on other modes).

In the following, the *Out categories* of a concept X are defined as those that X participates in (i.e. X = *dog* → *animal*), and the *In categories* of a category Y as those concepts that participate in Y (Y = *dog* ← *retriever*). Note that COGBASE does not distinguish programmatically or theoretically between concepts and categories; the two are expected to blend into and cross-constitute one another. Thus, any such distinctions made here are strictly expository.

In semantic atoms, the starting concept is referred to as the FROM concept and the end concept as the TO (i.e. *FROM* $\overset{primitive}{\longrightarrow}$ *TO*).

Below, the semantic atom X → *FACILITATE* → Y indicates that X can often be used to achieve the state of the world described by Y.

Examples:

vocal cord → *FACILITATE* → *sing*

hammer → *FACILITATE* → *build*

The atom X → *GOAL_CHANGE* → Y indicates that when X is encountered, people often change their immediate goal to Y.

Examples:

hungry → *GOAL_CHANGE* → *eat*

see money → *GOAL_CHANGE* → *pick up money*

X → *CONCEPT_ASSOC_CONSTITUENT* → Y indicates that X is loosely associated with being part of Y. X may not always be part of Y, but it is often enough so that it is worth noting.

Examples:

heating element → *CONCEPT_ASSOC_CONSTITUENT* → *heater*

engine → *CONCEPT_ASSOC_CONSTITUENT* → *car*

Primitives beginning with *T-* are *temporal* in nature, with *T-0* atoms, for example, indicating process prerequisites (i.e. FUEL is required for a FIRE), *T-1* primitives contributing information about initial process stages, and *T-DURING* primitives indicating information relevant as processes advance. In the algorithms below, the notation $\overset{+}{\longleftarrow}$ denotes addition assignment (+ =).

[2] Drawn from the Twenty Newsgroups dataset: http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html.

11.1. User additional concept interests

First, we consider the User Additional Concept Interests algorithm (sample results provided in Section 10.3).

Data: Input Concept, Use In/Out Categories (**bool**),
Include Sort Score in Sorting (**bool**),
Include Confidence Score in Sorting (**bool**)
Result: Augmented Additional User Concept Interests
UseConcepts $\longleftarrow$ *Input Concept*;
if *Use In Categories is True or Use Out Categories is True* **then**
RawCats $\longleftarrow$ Retrieve In/Out Categories of *InputConcept* ;
FilteredCats $\longleftarrow$ $x \in RawCats$ such that category node degree $\geq$ min
(there must be minimal data for each to allow noise filtering, and extremely sparse concepts are likely noise);
UseConcepts $\overset{+}{\longleftarrow}$ FilteredCats;
end
CollectedData $\longleftarrow$ {};
for $c \in$ *UseConcepts* **do**
CollectedData $\overset{+}{\longleftarrow}$ all TO concepts for atoms of specific primitives (outbound FACILITATE, inbound GOAL_CHANGE, inbound CONCEPT_ASSOC_CONSTITUENT, others) where FROM = concept c;
end
CollectedData $\overset{+}{\longleftarrow}$ all inbound nodes for c;
FinalData $\longleftarrow$ $c|c \in CollectedData, count(c) > threshold$;
(where $count(c)$ is the number of times c appears in *CollectedData*);
OutputAugmentation $\longleftarrow$ $\{SortScore(c), ConfScore(c)|c \in FinalData\}$;
$\mathcal{O} \longleftarrow \{sort(FinalData), OutputAugmentation\}$;

11.2. User goal inference

Next, we examine User Goal Inference (Section 10.2).

Data: Input Concepts Vector
Result: Goal Vector $\mathcal{O}$
RetrievedData $\longleftarrow$map(retrieve following primitives for c: inbound GOAL_CHANGE, INCREASED_LIKELIHOOD_OF, outbound T-0, T-1, T-LAST, T-DURING, FACILITATE) over Input Concepts Vector;
$\mathcal{O} \longleftarrow \cap_{S \in RetrievedData} S$;

12. Accessing COGBASE

In order to provide a means for scholars and practitioners to try COGBASE for themselves, a permanent API and documentation set can be found at: http://cogview.com/cogbase. The author warmly welcomes questions and comments regarding specific COGBASE applications.

13. Conclusion

COGBASE demonstrates how commonsense and other forms of nuanced knowledge can be theorized, stored, reasoned over, and made available in the form of a *semantic prior* for use in machine learning and reasoning systems.

COGBASE supports manifold new possibilities for taking semantics into account within AI, Big Data, NLP, NLU, and more, providing a core theory and knowledge base for nuanced, high surface-area multimodal reasoning and representation.

The system presented here provides powerful tools for extracting information from social data, implicit knowledge, complex contexts, difficult-to-parse syntax, and affective content.

Anticipated future work includes feature-based opinion mining (together with COGPARSE), metaphor processing, and further demonstrations of the precision and recall improvements made possible by the robust combination of semantics with machine learning techniques.

Acknowledgments

The author would like to sincerely thank the Carnegie Mellon University Language Technologies and Robotics Institutes, the United States Department of Defense (including DARPA contract numbers NBCHD030010 and FA8750-07-D-0185), the Singapore Ministry of Defence, and Singapore DSO for providing funding for this work.

Disclaimer: The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing official policies or endorsements, either expressed or implied, of DARPA, DSO, the U.S. or Singaporean governments, or any other sponsors.

References

Breslin, J., & Decker, S. (2007). The future of social networks on the Internet: the need for semantics. *IEEE Internet Computing*, *11*(6), 86–90.
Brooks, R. (1999). *Cambrian intelligence*. MIT Press.
Cambria, E., Olsher, D., & Rajagopal, D. (2014). SenticNet 3: a common and commonsense knowledge base for cognition-driven sentiment analysis. In *Proceedings, AAAI*.
Cambria, E., Rajagopal, D., Olsher, D., & Das, D. (2013). Big social data analysis. In R. Akerkar (Ed.), *Big data computing* (pp. 401–414). Taylor & Francis (Chapter 13).
Cambria, E., & White, B. (2014). Jumping NLP curves: a review of natural language processing research. *IEEE Computational Intelligence Magazine*, *9*(2), 48–57.
Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *Proceedings, AAAI*.
Dreyfus, H. (1992). *What computers still can't do: a critique of artificial reason*. MIT Press.
Evans, V., Bergen, B. K., & Zinken, J. (2007). The cognitive linguistics enterprise: an overview. In V. Evans, B. K. Bergen, & J. Zinken (Eds.), *The cognitive linguistics reader* (pp. 2–36). Equinox.
Lakoff, G. (1990). *Women, fire, and dangerous things: what categories reveal about the mind*. University of Chicago Press.
Lakoff, G., & Núñez, R. (2000). *Where mathematics comes from: how the embodied mind brings mathematics into being*. New York: Basic Books.
Langacker, R. (1999). *Foundations of cognitive grammar: theoretical prerequisites. Vol. I*. Stanford University Press.
Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., et al. (2014). DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*.
Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT Technology Journal*, *22*(4), 211–226.
Manovich, L. (2011). Trending: the promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (p. 460). University of Minnesota Press.
McCarthy, J., & Hayes, P. J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In *Machine intelligence* (pp. 463–502). Edinburgh University Press.
McDermott, D. (1987). A critique of pure reason. *Computational Intelligence*, *3*(1), 151–160.
Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11).
Mueller, E. (2006). *Commonsense reasoning*. Amsterdam, Boston: Elsevier, Morgan Kaufmann.
Murdoch, T. B., & Detsky, A. (2013). The inevitable application of big data to health care. *JAMA*, *309*(13), 1351–1352.
Olsher, D. (2012a). Changing discriminatory norms using models of conceptually-mediated cognition and cultural worldviews. In Proceedings, CogSci. (pp. 2138–2143).
Olsher, D. (2012b). COGPARSE: brain-inspired knowledge-driven full semantics parsing: radical construction grammar, categories, knowledge-based parsing & representation. In *LNCS: Vol. 7366. Advances in brain inspired cognitive systems* (p. 1). Springer.
Olsher, D. (2012c). Full spectrum opinion mining: integrating domain, syntactic and lexical knowledge. In *Proceedings, ICDM* (pp. 693–700).

Olsher, D. (2013). COGVIEW & INTELNET: nuanced energy-based knowledge representation and integrated cognitive-conceptual framework for realistic culture, values, and concept-affected systems simulation. In *Proceedings, IEEE Symposium Series on Computational Intelligence* (pp. 82–91).
Olsher, D., & Toh, H. G. (2013). Novel methods for energy-based cultural modeling and simulation: why eight is great in Chinese culture. In *Proceedings, IEEE Symposium Series on Computational Intelligence* (pp. 74–81).
Pfeifer, R., & Bongard, J. (2006). *How the body shapes the way we think: a new view of intelligence*. MIT Press.
Rajagopal, D., Olsher, D., Cambria, E., & Kwok, K. (2013). Commonsense-based topic modeling. In *Proceedings, ACM KDD*.
Shanahan, M. (1997). *Solving a mathematical investigation of the common sense law of inertia*. MIT Press.
Sun, R. (1994). *Integrating rules and connectionism for robust commonsense reasoning*. New York: Wiley.
Todorova, Y. M. (2006). *Representing commonsense knowledge using answer set programming*. Universidad de las Américas Puebla.
Waskan, J. A. (2003). Intrinsic cognitive models. *Cognitive Science*, *27*(2), 259–283.