NSF Workshop
Research Challenges and Opportunities in Knowledge Representation
February 7-8, 2013
Arlington, Virginia

Position statements from participants
Challenges for Knowledge Representation: Integration with Natural Language

James Allen, University of Rochester

While most researchers would agree that building systems that can accomplish deep language understanding, i.e., identifying meaning, intention and entailments from natural language, we have made little progress towards developing such systems over the past few decades. Deep understanding will require extensive amounts of commonsense knowledge and an integration of language understanding with reasoning systems.

I suggest a key challenge for knowledge representation in the next decade is to create KR&R systems that significantly support deep language understanding. Such systems would enable dramatically better models for intelligent human-machine interaction, intelligent search, assistive systems and tutoring systems. While natural language research has focussed almost exclusively on structural models of language for some time (i.e., the so-called shallow models), many future applications require addressing key issues that have been considered more in the area of knowledge representation than NLP. At the moment, however, the gulf between the fields is large, and there are few attempts to bridge between the disciplines.

As an example of the divide between the fields, and why each field needs each other, consider a common problem both face, namely massing human-scale commonsense knowledge about the world. For NLP, this is needed in order to accomplish better disambiguation of meaning and intention, and to find inferentially connections between sentences for effective discourse understanding (i.e., understanding a text, or an ongoing conversation, as opposed to a single sentence). For AI in general, it is needed to enable intelligent agents that can reason about the human world (e.g., for more effective and useful robots), or intelligent assistants for design, planning or health care. There have been attempts to hand-code considerable amounts of commonsense knowledge (most notably the Cyc project), but the Cyc knowledge base has not found much traction in applications, either in supporting NLP or in building intelligent agents. Part of the problem is coverage gaps, but a larger issue is the lack of fit of the knowledge to applications. Especially with regard to NLP, Cyc’s logical ontology differs significantly from the lexically-based ontologies used in NLP. Rather, WordNet is by far the most commonly used resource in NLP for reasoning about entailments, mainly because of its coverage and because the knowledge encoded in terms of word meaning and relations between words.

If we truly want to attempt to achieve human-levels of commonsense knowledge, then much of that knowledge is going to have to be acquired automatically. And by far the most promising avenue for acquiring knowledge automatically is learning by reading. And to accomplish this, we need to make progress on unifying natural language semantics with knowledge representations.

Here are three of the most significant challenges:

Semantic Models -- we need representations that reflect the underlying semantics of words and the compositional structure of meaning in natural language

Scale - We need reasoning systems that can handle human-scale commonsense knowledge bases

New reasoning methods -- we need probabilistic/abductive/defeasible reasoning techniques that support key processes in NLP such as disambiguation and entailment, as well as allowing and reasoning with vague knowledge.
1 How KR Can Help Machine Learning

Knowledge representation and reasoning is devoted to the design, analysis, and implementation of inference algorithms and data structures. Work in KR&R has deep roots in reality: reasoning problems arise naturally in many applications that interact with the world – common-sense query answering, diagnosis problem solving, planning, reasoning about knowledge in the sciences, natural language processing, and multi-agent control, to name a few. Aside from their obvious practical significance, reasoning algorithms and knowledge representations form the foundations for theoretical investigations into human-level AI.

Machine Learning is a sub-field of artificial intelligence concerned with the computerized automatic learning from data of patterns. The intention of machine learning is to use some training data to detect patterns, later using those learned patterns to automatically answer questions and autonomously make decisions and execute them. Examples of machine learning are the models learned by computers to predict users’ preferences for books, TV shows, and purchasing decisions in grocery stores. There, training data is books that people have chosen in the past, and characteristics of those books and the people who chose them. Models learned from these training data are then used to predict likely other books that people would buy.

Probabilistic graphical and relational models are close to statistical machine learning and serve as a medium between machine learning and knowledge representation. Nonetheless, Machine Learning as a field finds such structures hard to use and does not know how to use many other structures and lessons from KR. Examples include high-level structures such as Kripke Models, modal logics, and quantifiers. Most importantly, many insights, practices, and results from KR do not translate easily into ML practice.

It is the charge of the KR community to present its results in ways that can be easily used by ML practitioners without much familiarity with KR. I present one such example potential development for KR here.

2 Machine Learning using Knowledge Representation (MLKR)

Present research on probabilistic representation, learning, and reasoning focuses on issues that involve large numbers of variables (large here is larger than, say, 100 variables). The difficult elements there come from the different ways in which distributions behave; Joint distributions over more than 100 variables may seem very different when in fact (from a human perspective) they seem almost identical. Humans make assumptions such as independence of random variables that do not hold in reality, leading to an incorrect perception of similarity of situations.

Machine Learning needs Knowledge Representation to make better sense of dimensionality reduction techniques and dimensionality-increasing techniques. Here are two examples of possible MLKR challenges:

Support Vector Machines with a Logical-Formula Kernel

SVMs use Kernels to project classification problems into exponentially larger spaces, enabling linear separators to classify properly in these higher-dimensional spaces. The projection to higher-dimensional spaces is done using combinations and comparisons of features in the feature vectors to be classified, separated. Logical formulas can
specify comparisons and closeness between concepts, and should therefore be used as informed kernels
to create classifiers that have superior precision.

**Principal Component Analysis relative to a Logical Formula or a Logical-Formula Basis** PCA is
a technique used to reduce dimensionality of a set of vectors so that only those components that are
more significant play a role, hence yielding robust classifiers and predictors. Just like linear algebra
where numerical vectors play the role of basis for the result of the PCA, Logical Formulas can serve
as a logical basis for interpretation of more complex concepts. Developing PCA-like techniques based
on logical formulas could create classifiers that can take combinatorial structures and background
knowledge into consideration efficiently and precisely.

### 2.1 Combining Logic and Probabilities

Many applications have both stochastic and non-stochastic elements. For example, robot control can include
high-level specifications in logic and a lower-level probabilistic sensing model. Also, Natural Language
Processing wishes to apply high-level knowledge in logic with lower-level probabilistic models of text
and spoken signals. Finally, many databases are logic based (e.g. an entry $\langle Eyal, Shavit \rangle$ is in database $fatherOf$
indicates the logical statement $fatherOf(Eyal, Shavit)$), while relationships between those databases and
recent extensions to databases (e.g., an entry $\langle John, Mary \rangle$ is an uncertain entry in database $loves$ with
probability 0.7).

Since 1990 there have seen much work in the AI community and the Databases community on the combi-
nation of logical and probabilistic expressivity. These works present languages that can express probability
distributions together with explicit references to objects, functions, and relationships, as in First-Order Logic
(FOL) (e.g., (?)). These languages are useful frameworks for many machine learning applications, and recent
works also show that they are useful for computational efficiency of inference (?, ?).

Research on the combination of logic and probability is ongoing. Current challenges include (a) applying
relational structure in speeding up inference and treating probabilistic models over many objects, (b) combin-
ing knowledge bases that are given already in probabilistic or logical form, and (c) extending representation
languages to include functions and equality of objects in sound and simple ways.
Processing knowledge, ranging from acquiring it to reasoning and solving problems using it, is central to the development of intelligent artifacts. This has been realized from the early days of Artificial Intelligence and the quest for suitable knowledge representation formalisms has been on since then. (One can of course trace the history of logic and problem solving to ancient times.) Since then significant progress has been made in the development of knowledge representation (KR) formalisms and logics and several systems are now available that can be used to represent knowledge and reason with them. However, with the exception of certain types of probabilistic knowledge there remains a big bottleneck with respect to knowledge acquisition.

Currently there are several points of tension between knowledge acquisition and logic based knowledge representation languages. Most logic based knowledge representation languages assume that the representation is done by human. This causes a big bottleneck because the number of humans who have learned the skill to represent knowledge in specific KR languages is small. A scalable approach to knowledge acquisition could be to develop ways to automatically obtain knowledge from text. While there is a significant body of research on information extraction from text, the extracted units are often relational facts. I.e., there has been very little research in automatically extracting more general forms of knowledge such as rules, defaults and constraints from text. We have made efforts in that direction [1,2] and have built a learning based system that learns to translate knowledge expressed in English text to representations in KR languages such as Answer Set Programming. Two additional dimensions of knowledge representation (KR) came to fore in such attempts. The first dimension is the suitability of an ontology and a given KR language in terms of allowing automatic translation from natural language text to representations in that language. The second dimension is the suitability of the ontology and the KR language in terms of a system being able to learn how to represent knowledge using that ontology and that language. We noticed that ontologies that are suitable for human representation of knowledge are not as suitable when the representation is done by an automated natural language translation system, especially when the automated system is a learning based system.

Another tension point is that many existing large knowledge bases use knowledge representation formalisms that do not have a fully developed declarative semantics. Examples of this include various frame based knowledge bases and CYC. Often times this has happened because of a distributed approach to knowledge acquisition. For example, the notion of cloning (where objects selectively clone properties from other objects, especially prototypes) in
some frame based systems is very helpful in distributed knowledge acquisition, but until recently did not have a fully developed declarative semantics. We have made some recent effort [3] to give declarative characterization of such useful constructs. But there needs to be more cohesion between the development of KR languages and the languages and constructs used in the creation of large knowledge bases.

A related aspect is that there is now a growing body of knowledge bases, that are partly automatically obtained and partly obtained in a distributed manner, that have not been taken advantage of in knowledge representation and reasoning systems and that have issues such as inconsistency and incompleteness. This includes linguistic knowledge bases such as Wordnet, Verbnet, Propbank, and Framenet and world knowledge bases such as Conceptnet, Google Knowledge Graph, DBpedia, Freebase, and SUMO. Using these knowledge bases in practical applications and overcoming any inconsistency and incompleteness issues is a challenge. A related challenge is developing question answering (QA) systems using these knowledge bases and especially addressing Why and How questions [4], which has not been addressed much by the broader QA community.

Looking further ahead on knowledge acquisition, one can borrow ideas and learn from language development with respect to computer programming. One possible lesson is to realize that human knowledge representation can be faster if done using higher level constructs; such higher level representation can then be automatically translated to lower level representation. This is practiced to some extent by knowledge base developers but needs to be looked at by the broader KR community.


Ontologies store application domain knowledge: relevant terms for classes and properties, relations between these terms, and description of instances of classes and their relationships. They are now widely used in the Semantic Web, bio-health, and engineering applications, to fix a shared terminology and background knowledge, and to help with data integration and retrieval. Standardisation of languages (for example OWL) has led to increased infrastructure and tool development efforts which, in turn, is leading to increased ontology development and usage. Advances in reasoning algorithms have allowed the development of services supporting tasks such as consistency checking, classification. The deployment and use of ontologies with applications requires the development and generation of those representations. Thus ontologies are not just about machine processing – there is a key role for users to play in the creation and maintenance of content.

Ontology engineering includes all aspects of ontology design, maintenance, and re-use, and often involves considerable comprehension effort for engineers to understand how an ontology describes and relates its terms. Various paradigms play central roles: e.g., reusing existing ontologies or selecting suitable parts to reuse. Maintenance similarly requires interaction at various levels in order to compare and understand different versions of ontologies. All these tasks require suitable support through methodologies and ultimately, tooling. Tools do exist for the support of ontology design and engineering (e.g. Protege), but there are still gaps in the functionalities required to support the ontology engineering process and tool suites lack maturity. Integrated development environments for ontology engineering that come with sound ontology engineering support encompassing reasoning, versioning, collaboration, modularity, the recording of provenance information (to support trust) and maintenance will help users in following suitable methodologies or best practice.

Existing tools – in particular editing tools – are, in general, aimed at those with some expertise in representation (as opposed to those with solely domain expertise). If we consider deployment and use, recent successes have been made with tools adopting “semantics by stealth”, using familiar paradigms (such as the spreadsheet) to support domain users in using ontology-based representations for annotation and content creation. Such approaches will be important in promoting wider spread use of ontology-based representations in applications.
From the early days on knowledge representation and reasoning for the control of robotic agents has been a key motivating topic if not the holy grail of Artificial Intelligence research. The views put forward in the Shakey project at Stanford have strongly influenced research in knowledge representation, reasoning, planning, and plan execution. Important characteristics of AI based robot control include that plans are sequences of actions that are to achieve a given goal state and that the robot reasons about which actions are to be executed in which order and abstracts away from how the actions are to be executed.

In recent years we have seen a number of robotic agents performing human-scale everyday manipulation activities such as cleaning an apartment, making salad, popcorn, and pancakes, folding towels, etc. Having realized such systems it is time to reconsider the role of knowledge representation and reasoning and the methods to be applied.

It is obvious that robotic agents cannot perform everyday activities as vague as “clean up”, “set the table”, and “prepare a meal” without comprehensive knowledge processing capabilities. To have the desired impact we have to investigate and develop knowledge processing methods that, given a vague task, are capable of inferring the information needed to do the appropriate action to the appropriate object with the appropriate tools in the appropriate way. If robotic agents are to be that competent their reasoning must not stop at actions such as pick up an object. Even for a simple action such as pick up an object, the robot has to decide where to stand, which hand(s) to use, how to reach for the object, which grasp type to apply, where to place the fingers, how much grasp force to apply, how much lift force to apply, etc. These decisions are context and task dependent. How to grasp a bottle might depend on whether I want to fill a glass with it or whether I want to put it away. How to grasp a glass might depend on whether or not it is filled. If knowledge processing does not reason about these aspects of robot activity it misses great opportunities for having substantial impact on robot performance.

I believe that in order to build more capable knowledge processing capabilities for robotic agents we have to bridge the gap between knowledge representation and processing and other parts of the robot control system. The data structures that the robot uses to produce actions should be considered as a virtual knowledge base. Thus when the knowledge processing system is asked for the position of the robot this position should be looked up in the low-level data structures of the robot. The abstract predicate of reachability could be determined by the existence of an inverse kinematics solution for a holding position for the object. Whether a candidate position for placing an object is stable could be computed by checking whether object would move in a physics
simulation of the observed scene when placed at the respective position. Or, visibility of objects can be approximated as the fraction of the object that is determined as visible through offscreen rendering methods.

The knowledge-enabled control of robotic agents in everyday manipulation tasks will be a great opportunity for knowledge representation and reasoning research to investigate the longstanding research challenges regarding common knowledge, naive physics and commonsense reasoning.
The abundance of inexpensive preference data facilitated by online commerce, search, recommender systems, and social networks has the potential to stretch the boundaries of both (individual) decision support systems and (group-oriented) social choice. Specifically, concepts and models usually applied to high stakes social choice domains such as political elections, public or corporate policy decisions, and the like, will increasingly find themselves used in the lower stakes, high-frequency domains addressed by online systems. These present interesting challenges and opportunities for knowledge representation and reasoning schemes that support parsimonious representation, effective inference (both exact and approximate), and data-efficient learning. Current representation schemes for preference data can be roughly divided into two categories:

- Simple, primitive representations, often lacking a clear semantics, but designed to support fast recommendation and learning. One example is simple score-based schemes for collaborative filtering that support very effective learning and inference (e.g., using techniques like probabilistic matrix factorization). Multiattribute extensions often assume simple linear utility models, again without clear semantics.

- More elaborate schemes for preferences over multiattribute data, with semantics based on standard utility-theoretic notions (such as preferential independence). Examples include CP-networks (for qualitative preferences) or its extensions. While having firm semantics, effective inference is often unattainable, and relatively little attention has been paid to approximate inference or learning such models from data.

In some models, preferences are represented in monetary terms (e.g., willingness-to-pay or bid values), in which case semantic issues are less critical, but effective representations are still needed. Representations for bids/preferences in combinatorial auctions are one example.

The difficulties facing preference representation in some sense parallel those that have been tackled (in many cases, successfully) in probabilistic inference. Effective graphical models, relational and first-order representations, and fast inference and learning techniques have made the use of sophisticated probabilistic models very much the standard in many data-rich domains. Similar advances are still needed in the area of preference representation. However, representation, inference and learning in preference models present some distinctive challenges. For example:

1. Unlike probabilistic models, in which one assumes some underlying "ground truth", preferences are personal and (potentially) very idiosyncratic. To what extent should representations (and learning and inference techniques) be designed to support such diversity of preferences?

2. Due to their idiosyncratic nature, the preferences of an individual may not be sufficiently predictable given data about their choices or those of "similar" individuals. Explicit preference elicitation—one might think of this as a personalized form of active learning—then becomes critical. The behavioral and psychometric assumptions one makes about an individual's ability to assess their preferences then plays a critical role in the design of our representations and learning/interaction techniques in a way that does not often arise in machine learning.

3. Aggregation of preferences is one of the most important algorithmic tasks facing researchers and practitioners (in recommender systems, in computational economics, in computational social choice, and a variety of other areas). Applications range from resource allocation, market design, "data-driven democracy" and computational voting, to marketing, advertising, and product line design. Unlike aggregation of multiple probabilistic/statistical models, the range of aggregation criteria (or social choice functions) is almost limitless. The design of representations that support a variety of forms of preference aggregation (both exact and approximate) should prove to be challenging, but very valuable.
Understanding Spatial and Spatio-temporal Data: from cameras and sensors to high level abstractions

Anthony G Cohn, University of Leeds, UK

Although there has been an increasing level of research into KR formalisms for representing and reasoning about spatial knowledge, much of this has been divorced from real data sets. I.e. the research has simply assumed that spatial information is readily available in symbolic form, and moreover, many formalisms assume that the data is clean and noise free. But the most readily available spatial data comes from cameras and from other spatially located sensors (e.g. GPS transceivers in mobile agents). There are many applications which require sophisticated understanding of such data, or can be usefully augmented with it, from surveillance, to mobile assistance, to environmental monitoring. The use of qualitative spatial representations naturally provides some relief from both the volume and noisiness of such data, but the problem of how best to abstract to a qualitative/symbolic representation has not really been properly addressed, nor has the problem of combining qualitative and quantitative representations and reasoning, or indeed different kinds of spatial knowledge (topological, orientation, size, distance…).

Understanding visual data has been a challenge for the Computer Vision community for decades, and whilst much progress has been made in methods which attempt to understand such data using continuous/numerical techniques it is only recently that interest has (re)started in trying to extract symbolic high level representations from video data. Since video data is inherently noisy (e.g. owing to changing lighting conditions) and the high variance in the presentation of activities visually, extracting symbolic hypotheses is highly challenging. The challenge is made that much harder by the sheer volume of data, both already “out there on the web” or acquired in real time; but on the other hand this sheer volume of Big Data also provides mitigation for the problem since there is often redundancy (e.g. through multiple kinds of sensors, or spatially overlapping sensors). Another form of mitigation can come in the form of background knowledge about how the world is, and how activities progress, so as to help understand missing data, correct noisy data, and to help integrate and fuse conflicting data. In turn, this brings the challenge of where such background knowledge comes from, and in particular whether it all has to be manually specified, or whether it can be automatically acquired, through data mining/machine learning techniques (which of course themselves face all the challenges already outlined – noisiness, abstraction, data volume…, as well as the problem of where supervision, if any, comes from).

Another challenge in this area is to combine data acquired from sensors with language data – consider for example a cookery programme, and the commentary from the chef and other people on screen and the visual images of the ingredients being prepared; there is some temporal synchronicity here, but it is not perfect; there is extra information in both data streams and “superfluous” information (e.g. where they first tasted this particular ingredient).

The temporal dimension in sensor data provides many challenges in its own right, ranging from synchronizing separate data streams which may not be globally clocked, to investigating the prediction, postdiction and gap filling problems which all have special interest when spatial data is concerned because of the continuity constraints that need to be imposed. But there are also problems of granularity to consider – activities happen over many different temporal scales, and understanding and representing these is important, again particular when trying to make sense of Big Data. In fact, the granularity issue applies to space too – depending on the spatial resolution, objects may or may not be “visible” and this impacts on whether activities and processes can be discerned too.
Towards Model-Based Thinking – Johan de Kleer (PARC)

Driving Application: How can we automate the correct design, diagnosis and operation of complex cyber physical systems? Such systems range from bicycles to modern airplanes. Right now there are 1000’s of automated tools such as Mathematica, MATLAB, Modelica, Simulink, SolidWorks, ... These tools help engineers design ever more complex cyber physical systems. However, each tool is an idiot savant, i.e., very good at doing one thing without being able to incorporate much of the context. They are extremely brittle by AI standards. Engineering solvers are only as good as the models it is supplied and suffer from mathematical instabilities. It is very hard to determine whether the output of an ODE simulation or a PDE simulation accurately describes the behavior of the actual physical system - is there error in the model or the tool? This deficit is only becoming worse as the cyber physical systems themselves incorporate complex software which models itself. We are not going to be rescued by Moore’s law as the rate of increase in supercomputing for design tools is similar to the increase in performance of their embedded software. Both are exponential.

Why now: The idiot savant tools seem to have hit a wall. Not surprisingly, given they are all point solutions suffering from garbage-in-garbage-out. We need to start looking at is the glue between them – the commonsense reasoning that designers tacitly use when they design systems with these tools. Engineers use models throughout. Mapping among such models are usually based on qualitative analysis, teleological analysis and analogy, not only inference. Also, I see more and more mechanical, aeronautical, civil, manufacturing engineers incorporating Knowledge Representation concepts (ranging from ontologies, non-monotonic reasoning through to Markov Logic) in their languages and tools. Often badly. Given the broad interests in the engineering communities in AI ideas, now is a unique opportunity for (1) us in the AI community to scale our ideas and leverage this larger community interested in using our ideas, (2) learning from the engineering community the challenges for KR.

Challenges:

1. On the short term: More of the more structured analysis of work in KR needs to be focused on engineering problems and embodying thinking in all phases of design. I see two ways to approach this: (a) more KR people can start to address the KR challenges that arise out of engineering tasks such as articulating requirements and building models, (b) some way to support the interaction of these two communities which usually do not interact.

2. In the longer term, I see the system engineering problems as presenting a very strong challenge for KR. Most KR approaches are built on a logical paradigm of facts, relations, rules, ... And the underlying mechanism are some form of logical inference. I believe extending current KR paradigms, or for that matter, current formal methods more generally, is a not the big opportunity in the long term. Let us reflect on what humans actual do due when they design, operate, troubleshoot, ... systems. They all operate with models. Now that is not saying much because engineering tools operate with models. The challenge is how these models co-exist. I was stunned to discover that complete models for any cyber physical system I know of simply do not exist. People are amazingly good at working with partial, approximate and incorrect models. We build one model for the dynamics. We build another model for synthesizing the control system. The relationships between such models are rarely written down. It is this model-based thinking that KR needs to be start working on.
NSF Workshop: Research Challenges and Opportunities in Knowledge Representation

Michel Dumontier, Associate Professor of Bioinformatics, Carleton University, Ottawa, Canada

Health Care and the Life Sciences (HCLS) have long been at the leading edge of applying advanced information technologies for the purpose of knowledge management, knowledge retrieval and knowledge discovery. In the past decade, the HCLS community has been investigating the use of Semantic Web technologies being developed under the auspices of the World Wide Web Consortium (W3C). Semantic Web technologies have emerged as vertically integrated stack of standards for formal knowledge representation and reasoning that are free for all to use and are subject to continuous improvement with input from academia, industry, government, and hobbyists. Our Bio2RDF project uses Semantic Web technologies to publish, integrate and query billions of biomedical statements, and we have recently demonstrated how these technologies can be used to, check the consistency of knowledge, and discover new facts using automated reasoning over integrated terminologies and use semantic web services to answer fairly sophisticated questions. Although an increasing number of efforts have demonstrated the promise of the Semantic Web as a framework for large-scale, distributed knowledge management for biomedical informatics, substantial investments must still be made to increase community adoption, promote standard practice and foster technological innovation.

There is little doubt that building an effective HCLS Semantic Web will be a long, hard and tedious process and it’s worth thinking about the major challenges that lie before us. First and foremost is how to manage the multiplicity in overlapping or conflicting terminologies as well as the heterogeneity in knowledge represented with RDF/OWL. While these not new KR problems, we think having honest discussions based on meaningful scientific investigations that examine the merits and drawbacks of differing representations is key to achieving progressively more effective representations. Part of the solution requires continued research in methods for evaluation, especially task based evaluation. In the interim, we believe that it will be important for us to reconcile the different representations. While one approach will involve making it easier to generate mappings from one dataset to another, a second and longer term approach is to foster grass-roots approaches to capturing their needs. With adequate documentation, users should be able to come up with lightweight specifications that capture the majority of their needs, while learning about the benefits that formal knowledge representation can offer them.

Another major opportunity lies in bringing software onto the emerging data-centric Semantic Web. We believe that only a set of simple guidelines (e.g. SADI framework) are sufficient to unleash the vast potential of new information that can be automatically discovered through the invocation of a semantic web service. Thus, functional bits of software may then become part of other software and relevant data can be automatically computed as evidence for some question of scientific interest. Taken together, an increased understanding of how to effective represent knowledge based on scientific research, mapping technology to interconvert different knowledge representations, along with access to dynamic services, will provide a rich semantic web under which to pursue biomedical knowledge discovery.
Knowledge Representation and Reasoning, 2020
Tim Finin, University of Maryland, Baltimore County

The current decade presents challenges and opportunities for knowledge representation and reasoning -- what should we hope for by its end? Psychologists and our own intuitions suggest that people use many kinds of representation and reasoning techniques and can integrate their results to understand the world, make decisions and act. Daniel Kahneman’s popular book *Thinking, Fast and Slow* has many examples.

Finding a way to support this integrated diversity in machines is a key problem that remains largely unsolved. Partly this is due to the way we do research, which tends to concentrate on one paradigm at a time -- so we have oscillated between focusing on numerical and logical approaches several times during my professional life. Advances in tools and data availability has enabled or revived approaches, as in the recent explosion of data-driven techniques. Finally, new potential applications often arrive with needs that, if met, will yield great value, as developing better Web search systems did in the last decade.

My current knowledge representation interests are focused on supporting information extraction from text, where KR&R systems can play many roles. These include (1) defining an organizing ontology of concepts and relations with associated constraints and rules; (2) representing background knowledge of entities and facts relevant to the application or context; (3) supporting metadata such as provenance and certainty; (4) providing reasoning modules of various types; and (5) implementing a software framework supporting knowledge bases that can be used in real software systems.

Here is a short list of capabilities that I would like in future KR&R frameworks. While none are completely novel, I hope that they would be supported by open standards like those developed by the W3C for the Semantic Web and also available in software systems (some open source) designed for interoperability and (where possible) scalability.

- A flexible integrating framework for combining evidence from diverse sources, including deductive reasoning, procedures, machine learning components, graph analytics, etc.
- Systems for combining probabilistic and logical reasoning in natural and scalable ways
- Support for integrating or resolving conflicting information and identifying and dealing with noise, errors, lies, differing opinions, fluents, different levels of abstraction, etc.
- Tools for managing contexts, e.g., foregrounding concepts, entities, events, and relations that are relevant to the current situation, task or time
- Context mechanisms that allow one to tentatively add fact and draw conclusions while isolating the effects on the knowledge base so they can be easily undone
- Better support for temporally qualified data including knowledge about how frequent fluents change and how past values constrain or influence the likelihood of future ones
- Practical way to “age” provenance data, e.g., archiving or deleting it based on its age, importance and potential utility
- Property value and cardinality constraints that include expected probability distributions and not just types, enumerations or numerical ranges (for values) or ranges (for cardinality)
Formalizing human conceptual structure to support STEM education
Kenneth D. Forbus, Northwestern University

Driving application: Intelligent tutoring systems and learning environments incorporate formally represented models of the domain and skills to be learned. Such systems have already been shown to be valuable educationally in some domains, e.g. learning algebra. The potential for such systems to revolutionize education, by offering anytime, anywhere feedback has been recognized in prior NSF studies. A key bottleneck in such systems is the availability of domain knowledge. Moving beyond formal domains and into STEM learning more broadly will require new kinds of ITS', fueled by substantial knowledge bases that incorporate both specialist knowledge and commonsense knowledge. Commonsense knowledge is both important for interacting with people via natural language, and because many student misconceptions are based on everyday experience and analogies with systems encountered in everyday life.

Why now: Progress in artificial intelligence and cognitive science more broadly has led to a deeper understanding of how to represent aspects of human mental life, including events, causality, and explanations. For example, qualitative reasoning research has led to educational systems that have been used in a variety of domains, including thermodynamics, physics, and ecology. Such systems demonstrate that scaling up to a broader range of domains could provide broader impacts in terms of new educational systems. Moreover, progress in research on learning by reading is reaching the point that scaling up in terms of total amount of knowledge in systems is becoming possible (e.g. IBM’s PRISMATIC knowledge base, learned by reading, contains over 990M frames).

New challenges and questions:

1. Today’s reasoning systems work well in carefully engineered, narrow domains. By contrast, human reasoning is robust, flexible, and operates over broad domains of knowledge. Understanding how to mechanize human reasoning processes offers both the promise of deeper scientific understanding of human cognition and creating systems that can reason with scales of data that are beyond human processing.
2. Building up large commonsense knowledge bases, by reading, sketching, vision, robotics, and/or crowdsourcing.
3. Understanding tradeoffs in representation/reasoning effectiveness. IBM’s Watson showed that structured, relational representations led to factoid Q/A performance that far surpassed what was possible with purely statistical, word-based representations. Interestingly, Watson’s representations were also very shallow, encoding the contents of particular sentences at linguistic levels. By contrast, a Northwestern learning by reading experiment (text plus sketches) showed that deeper Cyc-based representations were useful for answering textbook problems. Interactive AI systems may need both types of representations, so understanding the tradeoffs between them becomes an important question.
The era of “big data” has arrived – we are awash in data, on the web, from scientific experiments, from sensor data and more. While there is a great deal of focus is how to store and manage the data, there is less understanding of how to transform the data into useful and actionable information. In part, I believe this is because not enough attention has been paid to knowledge representation and reasoning; in particular approaches that are able to combine statistical machine learning approaches with rich, structured knowledge representations. Machine learning methods need to be able to exploit structure in the data and express the data in some form of knowledge representation; knowledge representation methods need to be able to represent uncertainty that may be the result of some form of statistical inference; and reasoning methods need to be able to answer queries in manners that are not brittle, that express rankings or probabilities, that flexibly reason about only the relevant information, and that take user preferences and context into account.

Within the machine learning community, the area of statistical relational learning (SRL) [1] has emerged over the last decade as one approach that combines logical information (typically a restricted form of first-order logic) with probabilistic models (typically graphical models). There have been a series of workshops and tutorials over the years, and there is a small thriving SRL community. Recent work has focused on ‘lifted inference’, which is an approach for scaling inference. In coming years, I believe that ‘structure learning’ will be a focus area, and to properly do structure learning requires tight integration of knowledge representation into the search over models.

An important challenge is how these methods can be used to turn big data into useful answers. Much of what goes into this is the ability to align and match data from multiple sources (ontology alignment, schema mapping, entity resolution, etc.). In addition, other important problems include 1) inferring a knowledge graph from noisy extracted information (automatic knowledge base construction), 2) on-the-fly query-time knowledge representation alignment and construction (the ability to only align and reason about the portion of the information that is relevant to the query at hand) and 3) context-sensitive knowledge representation. The latter uses context and user preferences to allow one to select context-specific data fragments and supports ‘good-enough’ reasoning that is tractable (rather than trying to construct a single, globally consistent knowledge representation that may be inconsistent or computationally intractable to reason with). Another interesting related challenge is understanding when light-weight ontological information is sufficient for answering queries correctly versus when more sophisticated reasoning about consistency and other forms of logical reasoning is required.

Overall, I believe there are exciting opportunities for combining knowledge representation and reasoning with machine learning and reasoning under uncertainty, information extraction, and planning and search, and that these subfields of artificial intelligence have much to contribute to the big data movement.

My research focuses on intelligent systems that use knowledge representation and reasoning to assist people to do complex tasks. I have addressed a wide range of tasks: scientific data analysis, managing to-do lists, using data from the Web, etc. There are many challenging research areas, I highlight here a few topics that I find most interesting.

**Acquiring Knowledge from People**

We need intelligent systems that can acquire knowledge from people, whether new ways to do tasks or simply people’s preferences for how the system should behave. Acquiring knowledge directly from people will always be a necessary skill for intelligent systems, even if they are able to acquire much of their knowledge through machine learning approaches.

Key research questions in this area include: How can people extend the knowledge in a system? How can people understand what a system has learned on its own and help it to extend that knowledge? How can people correct misconceptions in a knowledge system? How can intelligent systems learn from several people who are providing overlapping information?

**Meta-Reasoning about Knowledge Limitations**

Intelligent systems used to assume that their knowledge was consistent and complete, assumptions typically made by logic reasoners. Intelligent systems must be able to operate despite containing knowledge with varying quality and coverage. They should be able to understand both their capabilities and their limitations as their knowledge grows over time. For example, a system that does not have enough knowledge about a topic should perhaps caution a user that its knowledge may not be sufficient to tackle a given task or answer a particular question.

Key research problems in this area include: What meta-reasoning capabilities does an intelligent system need in order to assess what it knows and what it does not know about? How can a system use the awareness of what it does not know and decide not to take certain actions based on its lack of knowledge? How can this awareness lead to seek new knowledge? Can intelligent systems find other systems that can complement their knowledge or capabilities?

**Trust and Provenance**

Intelligent systems used to be closed systems whose contents were carefully controlled by their designers. In contrast, new knowledge is often obtained from open sources on the Web. The information is often of unknown origins and there is
often no prior history with many of the sources that may be used to assess their reputation. Intelligent systems must be able to assess whether to trust a particular piece of information, but how? Representing the provenance of information is key to be able to track what sources were used and how they were combined, so that a trust determination can be made. As of December 2012, there is a new W3C standard for provenance on the Web called PROV, which will soon begin to be used ubiquitously on the Web.

Key research problems in this area include: How do we design algorithms that generate trust assessments based on provenance records? How can these algorithms cope with incomplete or inconsistent provenance records? How can these trust measures be incorporated by intelligent systems in their reasoning? How can trust and provenance be shown to users so they understand the reasoning of these intelligent systems?

**Reasoning about Complex Tasks**

Intelligent systems cannot just focus on question answering. They should have the ability to automate tasks for people, reasoning about the complexity of the different steps involved and what knowledge about objects in the world is required for those tasks. I am particularly interested in scientific discovery tasks, and understanding how develop intelligent systems that understand those tasks and can carry them out to assist scientists, for example designing experiments or carrying out complex analyses on collected data.

Research challenges include: How can complex tasks be represented? How can we reuse knowledge and capabilities for new similar tasks? How can knowledge be transformed so it can be used for different tasks by different kinds of reasoners?

**Collaborative Knowledge Discovery**

Bringing people of different skills and expertise together to look at a problem is a good recipe for innovative ideas. In many areas of science, the amalgamation of different kinds of expertise results in improved understanding and new discoveries. This becomes increasingly challenging as we study more complex phenomena, particularly with the advent of big data. A long-term goal would be the development of intelligent systems that could mediate among people who have different expertise, and facilitate the synthesis of new insights. Perhaps these systems could even contribute to the problem with their own knowledge and reasoning ability.
The biggest opportunity for knowledge representation (KR) is to enable query answering over heterogeneous sources. The success of the World Wide Web has led to a proliferation of data sources on a diverse range topics. However, there is no mechanism for treating these sources as a single integrated knowledge base. Such a tool could enable the concerned parent to find potential treatments for a child with a rare disease, a citizen to uncover government waste and propose effective cost-cutting measures, or a scientist to learn of results from another discipline that lead her to a paradigm shift in her own area.

There are a number of challenges. The first set of challenges are the same as those for “big data”: volume, velocity, and variety. The traditional study of knowledge representation and ontologies is a tool to combat variety. Ontologies can provide formal semantics and controlled vocabularies. However, ontologies in and of themselves do not solve the variety problem. As seen in Linked Data, as thousands of ontologies start to blossom, we need to provide mechanisms to discover axioms that interrelate these ontologies. Although there has been significant work on automated ontology alignment in recent years, most of this work focuses on simple alignments, and does not consider that in many real world applications, alignment requires intersection (to combine two concepts), negation (to exclude some things), and complex value mappings. A related problem is that of instance mappings, sometimes known as entity coreference or entity resolution. We need domain-independent algorithms that are capable of creating high quality instance mappings for hundreds of millions of instances. The issue of volume has been addressed to some extent by the Semantic Web community. However, we must set our sights on expressive reasoning with exabyte knowledge bases. Velocity is the issue that has received the least attention so far. Reasoning with streaming, graph-oriented data is a critical problem. In particular, reasoning across information that comes from different streams or that arrives at significantly different times can be quite difficult.

Another problem lies in user interaction. If there is a large volume of high variety data, it is unrealistic to expect users to know which ontological terms can be used in their queries. One solution lies in natural language processing to translate a user’s question into a formal query. A slightly more obtainable goal would be an interactive approach that involves multiple natural language inputs in response to prompts from the system. Yet a third way is novel visualization of the structure of the data. For example, Heflin’s research team has created an interface that combines aspects of tag clouds, faceted browsing, and set theory in order to provide easy navigation through the hundreds of thousands of classes and properties in Linked Data.

A third great challenge is uncertainty and the closely-related issue of trust. The Web has given knowledge representation renewed vigor, but to fulfill the promise, KR must address the messiness of the Web head-on. We need KR languages and methods that allow for knowledge of varying quality and we should explore both probabilistic and modal logics as possible solutions to this problem. We must recognize that on the Web, there is often no universal truth, but instead nationalism, politics, religion and personal preferences lead to irreconcilable differences. Thus, trust and certainty should be tailored to each user and his/her peer group. However, this user adaptation must be moderated so as not to amplify the already prominent trend of many users to seek out media that reinforces their current beliefs and tune out dissenting voices.
Knowledge Representation in the Big Data Age

Pascal Hitzler, Kno.e.sis Center, Wright State University

Position statement for the NSF Workshop on Research Challenges and Opportunities in KR
Arlington, VA, February 2013

Enormous amounts of data are now at our fingertips. It is produced with breathtaking speed, while at the same time it differs widely in formats, perspectives, usefulness, reliability. One of the foremost research challenges in Computer Science is how to make non-trivial use of this wealth of data, and it seems obvious that the development of impact methods and tools will require cooperative efforts which span many sub-disciplines of Computer Science, while working hand-in-hand with application areas. Facing this challenge, KR needs to redefine itself as an application-oriented discipline. The following are three corresponding challenges which seem to be of central importance.

Closing the gap between deduction and induction. Current KR methods work at their best if data and knowledge bases are hand-crafted or hand-curated and adhere to a single modeling perspective. But this approach does not scale up to Big Data requirements, neither in terms of volume nor in terms of noise present in the data, which may come from errors or from different modeling perspectives. At the same time, inductive methods like those based on statistical machine learning or data mining are good at filtering out noise, generally perform better under higher data volume, and are designed to detect higher-level features in data—however they lack deductive reasoning abilities and largely remain black boxes. A successful dealing with Big Data necessitates that we have the best of both worlds available, which means that we have to establish methods which bridge the gap between logic-based knowledge representation and inductive learning and mining methods. This includes the development of applications which cross between these fields, as well as novel theoretical underpinnings for such an integration.

Developing light-weight KR for education and practice. KR research has resulted in a plethora of different logics for knowledge modeling, and many of them are well-motivated by formal or informal considerations. However, the landscape of logics is unwieldy and difficult to navigate even for the expert, and even the KR community itself is essentially split into sub-communities which concern themselves only with certain aspects of the KR landscape. We require a consolidation of the field which maps out main approaches, preferably in a modular fashion, in such a way that KR methods become more easily accessible and easy to apply for the non-expert. This will include cross-paradigm studies, but also a boiling down of complicated logics to practical requirements, which are often orthogonal to what the available deep theories can deliver.

Facing the educational challenge. The entry barrier to KR is very high. Anticipating higher demands for knowledge engineers in the near future, we need to design courses and curricula which educate in key skills required for applying KR methods in practice. Sound theoretical underpinnings are required, but the focus of such an education needs to be on the transfer of KR methods to application scenarios, starting with a clear understanding of knowledge modeling in different KR paradigms, and exposure to application domains and their peculiarities. A key challenge for developing and delivering such courses and curricula is to provide bridges for a transfer of theoretical aspects of KR to practice-oriented applications.

References

Ontology Enhanced Information Systems

Ian Horrocks

Surprisingly(!), my focus is on ontologies and ontology enhanced information systems.

Until recently, the main focus of KR/ontology research has been the provision of ontology engineering tools and infrastructure, i.e., systems that support the development of (at least) coherent and (hopefully) high quality ontologies. Of particular interest has been the development of logic based tools that allow for checking consistency, computing a class hierarchy, running regression tests, etc. This research has been very successful, and the use of such tools has now become standard practice in almost all mainstream ontology development efforts. For instance, an OWL EL reasoner is now systematically used in the development and maintenance of the SNOMED ontology, which is an important component of the national health IS of the UK and more than a dozen other countries, and is also used by many US health service providers (such as Kaiser Permanente). Similarly, the OBO foundry servers (which host a suite of more than 100 "orthogonal interoperable reference ontologies" in the biomedical domain) now run OWL reasoners 24/7 in order to automatically check the integrity and interoperability of OBO ontologies.

Although the underlying reasoning problems that the above mentioned reasoners are solving are (very) hard (ranging from PTime-complete to 2NExpTime-complete), highly optimised modern systems are extremely effective — ontologies containing hundreds of thousands of classes can be fully checked and classified by the ELK reasoner in only a few seconds using standard desktop hardware. We can therefore consider this kind of reasoning to be "a solved problem" — systems are robust, and robustly scalable, and are in daily use in mainstream applications.

When deployed in applications, however, the ontologies developed using the above technologies are typically treated simply as class hierarchies, with the rich knowledge captured in the ontological being largely or completely ignored. This is because deployment in applications typically means dealing with large data sets. Large data sets present a completely different challenge, but also enormous opportunities: more effectively exploiting valuable data resources in both the private and public sectors will likely prove essential for the future competitiveness of our industries. For example, in the Norwegian oil and gas company Statoil, exploration decisions are based on an analysis of data from previous drilling operations, and the quality of this analysis is the single most important factor in determining the outcome of explorations worth billions of euros. However, the limitations of existing systems make it impossible to locate all relevant data, resulting in analyses being less precise and more error prone, and in highly skilled analysts spending more of their valuable time searching for data than analysing it. Statoil have estimated that more effective exploitation of this data could add billions of euros per year to the value of their North Sea oil and gas production.

"Ontology Enhanced Information Systems" (OEISs) have the potential to address many of the critical bottlenecks in scenarios such as the one described above, e.g., by (i) providing a unified and domain-centric view of the data, allowing users to formulate complex queries using familiar vocabularies and conceptualisations; (ii) answering queries over semi-structured data sources; and (iii) enriching query answers with implicit information. For example, a Statoil geologist might want to retrieve data on porosity and depth for stratigraphic units from the Mesozoic era along the North Sea border between the UK and Norway. A suitable ontology would not only allow an appropriate query to be formulated in exactly these natural (to a geologist) terms, but would also deal with issues such as the inconsistent naming schemes used in the UK and Norway, and the fact that Jurassic units are implicitly Mesozoic.

The challenge for KR/ontology research is that, in order to be successfully deployed in a wide range of application settings, an OEIS must support a rich ontology language capable of accurately modelling complex domains, and at the same time offer query answering scalability comparable to existing information systems. Regarding the former, this can be assumed to mean supporting OWL, an ontology language that has been standardised by the W3C, is already in widespread use, and using which numerous ontologies have already been developed — for example, there already exist detailed OWL ontologies covering the Oil and Gas, and Healthcare domains. Unfortunately, no scalable query answering procedures for OWL currently exist: those that can fully handle OWL are not scalable, and those that are scalable can fully handle only relatively weak subsets of OWL. If we can repeat the success that has been achieved w.r.t. conceptual reasoning, then we could lay the foundations for a new generation of information systems that will revolutionise our ability to exploit data resources.
The physical capabilities of robots have recently made huge strides. The actuators are reasonably safe and reliable and the sensing is sufficiently accurate to recognize known objects in somewhat complex arrangements. Of course, there is much to be done to improve physical robots, but we can no longer argue that bad hardware is responsible for the inability of robots to operate effectively in everyday environments.

So, the difficulties must lie in the software, and those difficulties lie not in implementation, but in conception and algorithms. There are three critical areas: methods for representing knowledge, methods for updating the robot’s internal knowledge representation based on percepts and actions (we’ll call this belief-state update), and methods for planning, execution, and execution monitoring (we’ll call this action selection). Additionally, we can consider the problem of learning, which is often distinguished from belief-state update because the knowledge being acquired or updated may be more abstract or variable over a longer time scale. For now, we will not make this distinction.

I am interested in building a robot that can perform every-day household tasks, such as tidying a room, putting away groceries, or even making a meal. It should do so in an environment that has not been modified (much) to accommodate it and that involves other (human) agents, so that there are, in AI planning terms, exogenous events. This means that the robot cannot depend on the state of the world remaining the same unless the robot makes a change.

There are three major challenges for operating in a domain of this kind: a mixed continuous-discrete state space, the dimensionality of which is unbounded, substantial uncertainty about the current state of the world and about the effects of actions, and very long time-horizons for planning. The first two combine to present a significant challenge for knowledge representation: representing uncertainty over state spaces of unbounded dimension.

Let’s consider a relatively simple example problem: finding the pickles in the back of the refrigerator. To do so, the robot will need to do some combination of moving objects (possibly removing them temporarily, or pushing them aside to get a better view and selecting view-points for look operations). All of these actions ultimately take place in real, continuous space, and must be selected based on the robot’s current belief state. To do a good job of this, the robot needs a representation of its belief about the current state of the refrigerator:

- What objects and types of objects are likely to occur in the refrigerator?
- Where are they likely to occur? May need to represent this distribution in terms of likely locations for classes of objects (e.g., bottles of beer in the fridge door) and/or in terms of which types of objects tend to be stored together (condiments on one shelf, beverages on another)
- How densely packed are the objects in the fridge?
- Is someone likely to have used the pickles recently (in which case they would be near the front) or to have eaten them all (in which case they are not to be found)?
• Does the person who asked for pickles prefer sweet or dill?

• What is the pose (position and orientation) of a particular object that the robot is contemplating picking up (and also the poses of the nearby objects, so that it can plan a safe path for moving the arm and grasping the object)? What is the uncertainty of that pose? (If the uncertainty is high, the robot will need to do further sensing using cameras or touch before it is safe to attempt to pick it up).

The robot also needs a representation of the effects of its actions on its belief state, including how likely an object is to be pushed or knocked over when it tries to pick it up, whether looking at an object from a particular angle is likely to result in a reliable detection, etc. This knowledge is used both in the process of belief-state update and for planning. This is a difficult problem as well, but in the rest of this paper, I will focus on representing information about the current state.

In order to highlight some of the issues in state representation for this domain, I will outline the representation we are currently using in an implemented system. It has innumerable difficulties, but it begins to make an attempt on the problem.

We have an explicit representation for objects whose existence we have evidence for, either because they were asserted to the robot to exist in the problem specification or because they have been perceived. Each object has a type (we currently have a fixed known universe of object types) and a pose. We represent distributions on these features using independent multinomial distributions on each of the types and a joint Gaussian distribution on the poses of all the objects. (There is an approximation that must be handled delicately here, of putting a distribution on the orientations as well as positions). The Gaussian itself is a severe approximation; we augment it with constraints that objects must be supported by other objects and that they may not inter-penetrate. This allows to generate samples from the distribution that satisfy physical constraints. Inference in this representation is done using an unscented Kalman filter, with the constraints applied post hoc.

We have an implicit representation for other objects: we can represent distributions over the types of objects that might be in a region of space and distributions over which types of objects tend to co-occur. Inference in this representation is done via MCMC.

Together, these representations support a reasonable strategy for searching for an instance of an object type. Currently, the high-level planning part of our system uses a relatively traditional symbolic formulation in terms of fluents (but where the arguments of the fluents are drawn from infinite sets, such as the set of poses or of regions of space). The fluents make assertions not about the state of the world, but about the robot’s uncertainty about the state of the world. So, for example, we have a fluent BVRelPose(o1, o2, eps, delta), where o1 and o2 are objects, epsilon is a probability and delta is a vector of values; the fluent asserts that, in the probability distribution on the random variable pose(o1)−1.pose(o2) (that is, roughly, the difference between the two poses), 1-eps of the probability mass is assigned to values less than delta. That is, it asserts that we know the relative poses of these two objects reasonably well. It is useful, for example, as a condition on the relationship between the robot and a target object before it attempts to pick up that object.

Some additional considerations are:

• Making multi-scale representational models: at some scale, I only want to think about something as “the spice cupboard”. Later, I may need to reason in detail about the objects inside. Similarly for bunches of grapes.

• Representing more physical constraints and understanding how they interact with positional distributions. I might want to represent that an object is contained in a bag, so that if the bag moves the object does, too.
• The belief-space fluents are always derived from the underlying continuous-valued belief-state representation, but in the future we will also want to represent uncertain information about color, size, and many other object properties.

• How else can we perform reasoning in these uncertain models? When can we take the probabilistic belief state, make a deterministic approximation, and do our reasoning there instead?

• How can we extend to non-rigid objects, like cloth and liquid?

• There is an idea that only some subset of the representation should be maintained in a kind of “working set”, to have the belief state updated about those objects frequently, and to kind of “swap out” the other objects until we need to think about them. How can we manage this? What kind of belief update would we do when we “swap in” an object that we had been ignoring for a while?
The knowledge representation (KR) community has developed sophisticated languages and ontologies for representing the knowledge in diverse subjects, yet the amount of data that is actually represented in a KR system continues to shrink as an overall percentage of data available. Consider just the growing body of data available as part of the Linked Open Data cloud [1]. While this information is published in RDF, much of the data is published in RDF using only the schema of the original data source so there is no useful semantic description of the data. While there are rich ontologies for some of the Linked Data sources, these are the exceptions and not the rule. Then there is the rest of the Web, which provides the majority of the available data. On the Web, the data and services are available in any of a variety of formats and there is no attempt to provide any semantic description of the data at all. The challenge and the opportunity are to bring the rich set of KR languages and ontologies to the vast amount of data that is available today.

Solving the knowledge / data representation gap will lead to huge advances in our ability to exploit diverse sources of knowledge. Consider the domain of biology where there are huge investments in research, equipment, and data collection. The ability to find and reuse data is extremely limited because it is a largely manual process to find, understand, and use data generated by other researchers. But if all of the data within this domain were published and described with respect to shared domain ontologies, then researchers could quickly discover relevant data sources and then exploit this knowledge to more effectively conduct their research.

Closing this gap requires developing new methods, tools, and incentives to represent the huge amount of data that is available today. First, we need methods to build semantic descriptions of the growing amount of data that is being produced. Given there is already a huge amount of data that lacks the semantic metadata that describes it, we need semi-automatic methods to support the semantic description of this legacy data [2,4,5,7,8,9]. Second, we also need methods that can quickly and easily (and perhaps automatically) transform data between alternative representations since different representations of the same data are often needed for different purposes [3]. Third, given that the amount of data is so vast and dispersed and the knowledge of what it contains is highly distributed, we need easy-to-use open-source tools that enable the users of the data sources to describe their own datasets. Finally, we need incentives in the form of an immediate return in the time and effort invested in publishing semantic descriptions of data to encourage the use of such tools. These incentives could be in the form of useful software tools that provide capabilities that are enabled by the semantic descriptions of the sources [6].

By bringing knowledge representation techniques and tools to the data and services that are already being published on the Web, we have the opportunity to start a revolution in representing, discovering, and exploiting the vast amount of data available today.
References


Role of KR in Natural Language Understanding and Synergic KR

Yuliya Lierler
University of Nebraska at Omaha

Research in natural language understanding (NLU) focuses on the design of systems that process information expressed in natural language and use it in reasoning. Key components of such systems are responsible for (a) transforming input in natural language (NL) into logic form – expression in a formal language capturing semantic information carried by the input; (b) performing required reasoning task on this logic form. Creation of advanced logic forms and methods for their construction given NL input can be seen as common challenges of the NLU and KR fields. Automated reasoning plays an important role in KR as its ultimate goal to advance both accumulation and processing of knowledge. In linguistics the focus is on development of logic forms that capture numerous phenomena displayed by languages, but automation is rarely of concern in this field. Work by Bos and Markert on the system nutcracker for the task of Recognizing Textual Entailment (RTE) is an exceptional example when a logic form language – discourse representation structures (DRS) – is used both for capturing and processing information in NL using FOL theorem provers and model builders. Yet only a subset of the DRS language compatible with FOL is utilized by nutcracker.

There are a couple of challenging questions. How similar logic form languages are to KR languages? Are current KR languages and respective automated reasoning tools applicable to logic form languages or, in other words, can they address the phenomena that natural language expressions encode? Developing systematic ways for constructing logic forms given arbitrary NL input is also a challenge. System boxer by Bos is a rather unique example of a tool that translates English sentences into FOL. It has limitations. One of them is related to the fact that FOL is inappropriate for capturing NL input. Another peculiar aspect of NL expressions is that each word carries independent information that has to be represented formally in order to process utterances. For example, sentence John walks suggests not only that John executes an action walk but also commonsense knowledge associated with action walk – an entity walking changes its location. Designing procedures that combine logic forms stemming from a NL input together with “background knowledge” associated with its words poses another challenge. Lexical databases wordnet, framenet, verbnet, propbank, nombank, and commonsense knowledge bases as conceptnet, knext provide wide scope of machine readable knowledge. If and how can they be used to advance NLU and KR? Recent tasks RTE, COPA (Choice of Plausible Alternatives), and Winograd Schema are good examples of problems that require sophisticated logic forms, background knowledge, inference mechanisms. These tasks may serve as an inspiration for the directions of research discussed above.

As KR matures, it offers a variety of languages and reasoning procedures appropriate for different tasks at hand, such as scheduling and planning, to name a few. For many application domains, including NLU, systems that combine multiple KR languages and tools are necessary. Developing systematic means for combining (a) heterogeneous KR languages and (b) various reasoning techniques under one roof by no means is a solved issue. Advances in satisfiability modulo theories and constraint answer set programming demonstrate a potential for this direction of research. For instance, constraint programming is known for being an efficient tool for solving scheduling problems, whereas answer set programming is effective technology for addressing elaborate planning domains. Constraint answer set programming that unifies these two KR sub-fields is best for solving problems that require both scheduling and planning capabilities of underlying tools.
Research Challenges
in Nonmonotonic Knowledge Representation

Vladimir Lifschitz, University of Texas

Early research on nonmonotonic aspects of knowledge representation was motivated by the need to formalize defaults (in particular, the commonsense law of inertia) and by the need to describe the declarative meaning of negation in Prolog. By mid-1990s it became clear that Reiter’s frame default provides an adequate formalization of the commonsense law of inertia; that the meaning of negation in Prolog can be explained in terms of default logic; and that default logic can be partially implemented using the computational ideas exploited in satisfiability solvers. This work has led to the creation of modern (“post-ADL”) action description languages, capable of representing actions with indirect effects, and to a closely related development—the emergence of answer set programming (ASP).

Here are some of the challenges related to these directions of research.

- **Modular action languages**: allowing an action description to consist of reusable modules; using a language of this kind to create a general-purpose database of commonsense knowledge about the effects of actions. The availability of such a database would facilitate describing large dynamic domains.

- **ASP with functions**: defining the semantics of rules that characterize functions directly, without encoding them by predicates; implementing ASP languages that incorporate such rules. This will simplify representing non-Boolean fluents in ASP and will make automated reasoning about actions more efficient.

- **Theoretical problems of ASP**: simplifying and clarifying the semantics of aggregates and other constructs that have been added to the basic language of logic programs with negation as failure. This will improve the methodology of developing provably correct ASP programs and provably correct answer set solvers.

- **Computational problems of ASP**: developing new optimization methods for ASP; incorporating computational methods of constraint programming and procedural implementations of functions. This will extend the class of practical computational problems that can be solved using ASP.
Knowledge Representation Challenges in the Next Decade: Making the Best of the Changing Landscape

Natasha Noy, Stanford Center for Biomedical Informatics Research, Stanford University

As any science, Knowledge Representation (KR) has to evolve as the environment and, in this case, the available knowledge to represent, evolves. I will focus on three fundamental changes in the last decade that affect KR: (1) the availability of linked open data; (2) the growing appreciation among scientists that they need some way to “tame” the data that they are producing; (3) the new paradigms of creating knowledge collectively.

Linked Open Data and what to do with it: We now have enormous collections of “real-world” structured knowledge in the form of linked open data—something that KR researchers never had the luxury of having. This data allows us to address new scalability challenges in reasoning and data integration. In many cases though, we treat LOD as a solution in search of a problem. We need to focus on real problems that having all this structured data can help us solve—either in the near future or after years of novel research. Working with data producers and consumers to understand the real problems that they have, will help the KR researchers not only to frame and address their problems in a way that is useful to others, but also will provide a framework for evaluating the efficiency of our approaches.

Big data science: Nothing provides bigger promise to KR than the need by scientists to process the growing amounts of data. Many of the methods that KR researchers have been working on will be critical in helping scientist “tame” this data: semantic annotation, data integration, reasoning. Collaborating with scientists to understand their problems in using data effectively will help us focus on the challenges and methods that will make real difference. So far, simple methods such as using class hierarchy in an ontology have been able to go very far in analyzing, for example, electronic medical records, or predicting adverse effects of drugs. Can we do better by using “more” knowledge or more advanced knowledge representation and reasoning?

Collective intelligence vs. KR: The way that the knowledge itself is created has changed dramatically in the past decade. We are now all contributing to knowledge creation, either knowingly or as a by-product of other tasks. This new way of knowledge creation requires new research on two major fronts: On the one hand, we need to understand better how to represent different views and opinions on the web scale and how to reason with contradictory and inconsistent knowledge coming from different users. On the other hand, this new paradigm of knowledge creation presents us with a unique opportunity to use human computation to address problems that have proved to be “AI hard.” We can explore whether we can use contributors from microtasks on Mechanical Turk, to citizen scientists on Zooniverse, to broad community of collaborators on ResearchGate and similar platforms, to solve problems of creating, verifying and evaluating formal knowledge bases and integrating heterogeneous data.
From the viewpoint of description logics and ontology languages, there are several unmet needs that adversely affect the use of large and sophisticated knowledge bases. The first unmet need is a lack of tools that aid developers, particularly developers who are not experts in modelling, in building and deploying large or sophisticated knowledge bases. There are tools for building ontologies, notably Protégé, but these tools are most suitable for experts and are designed for stand-alone use. What is needed are tools that are integrated into standard development environments, such as Eclipse, and provide all the usual facilities of these environments, including concurrent multi-user support with fine-grained control of rollback, display of changes between versions, and analysis of dependencies between parts of the knowledge base and between the knowledge base and the rest of the system. Much of this unmet need does not involve new research, just new development, but some of it, particularly ontology development for non-experts, requires research in human-computer interaction.

The second unmet need is a lack of sophisticated reasoners that can exploit the vast computational resources that have recently become easily available. Current performance reasoners for expressive ontology languages, e.g., HermiT, Pellet, and FaCT++, depend on very sophisticated central control of the reasoning process that attempts to do as little wasted work as possible. This central controller would have to be greatly modified to work well for multiple reasoning tasks, and then would also need to be decentralized lest it become a bottleneck.

The third unmet need is a lack of deployable reasoners for non-standard inferences, which are needed to support many kinds of applications. Some of these non-standard inferences, such as computing least common subsumers or explanation in ontology languages, have known algorithms, but these algorithms are not present in performance reasoners. Other non-standard inferences, such as abduction, have been the subject of some study, but their precise formulation is not certain in expressive ontology languages. Yet other non-standard inferences, such as abduction for query answering, have not yet been well-studied, even though they naturally show up in applications, for example in determining what would be needed to provide a known answer for a conjunctive query.

The fourth unmet need is a good way to bridge between description logics and ontology languages, with their open-world, unconstrained approach to information; external information sources that have a closed-world approach; and application code. There has been work in this area, with several notions of constraints in ontology languages bridging between open and closed worlds, and several good APIs for accessing ontology knowledge bases from programming languages, but none of the approaches are completely satisfactory. Certainly there is nothing similar to the close integration that can be provided between frame systems and programming languages.
Diversity of KR formalisms to enable multi-strategy reasoning/learning in Commonsense Reasoning

Doug Riecken - Columbia University Center for Computational Learning Systems

"to learn something well, you must learn it many different ways" - Marvin Minsky

Complex problems that require expressive problem solving strategies suggests: (1) the utility of gaining multiple perspectives of a given problem solving situation, (2) formulating and evolving one or more hypotheses of the situation, and (3) providing one or more explanations of each hypothesis. The need for a diversity of reasoning/learning techniques and knowledge representations when addressing complex problems continues to be an important research topic [McCarthy, 2002].

From my studies in commonsense reasoning, I am interested in complex problems spaces, such as situational awareness in social media environments, which require multi-strategy reasoning and learning techniques to effectively identify and respond to emerging situations of interest. One approach investigates the utility of independent reasoning agents, each agent being a distinct expert leveraging internal effective KRs best addressing the reasoning modality of the agent (e.g., spatial, temporal, causal, analogical, structural, functional, etc…); these agents collaborate in various groupings.

Specifically, I have been investigating formal representations for classes of problems that fit and are tractable to leverage Decentralized Partially Observable Markov Decision Processes (DEC-POMDPs) to enable multi-strategy reasoning and learning by collaborating sets of agents. Further, I am examining formal representations that will allow extending the complexity and scale of various problem spaces that our reasoning/learning techniques address.

Prior work on the M system [Riecken, 1994] provides an example of collaborating agents that generate multiple perspectives, hypotheses, and explanations of a problem space. Each agent is a complex expert in a specific reasoning modality (e.g., temporal, causal, etc…). M demonstrated an expressive performance of its agents collaborating via a centralized blackboard control mechanism. As a subsequent example, the RESIN system [Yue, 2009] highlighted extending the M work by applying a centralized Markov Decision Process (MDP) [Bertesekas, 2006] to implement RESIN’s blackboard control mechanism. This work lays the groundwork to advance the use of MDPs as an approach towards a decentralized control mechanism that would enable each agent with its own ability to manage collaboration with other agents versus a centralized controlling architecture.

More recently, the SNARE system [Riecken 2012] examined utilization of a diversity of knowledge representations to support a large number of distinct modal reasoning processes/agents addressing various commonsense reasoning research agendas. Collaboration and competition between various dynamic sets of agents is achieved via use of Decentralized Partially Observable Markov Decision Processes (DEC-POMDPs). The SNARE architecture leverages decentralized collaborating agents as an effective approach to: (1) reason from multiple perspectives, (2) provide a means by which each agent manages it reasoning and learning in collaboration with other agents, and (3) create a dynamic technique for agents to join and exit “agent communities”.

Several useful areas of study include: (1) KRs for each specialized modal reasoning/learning agent, (2) KRs to support dynamic communication of knowledge between groups of diverse agents, and (3) KRs and algorithmic treatments, such as DEC-POMDPs, to enable dynamic formation of agent collaborations.

References


Mark Schildhauer

There have been major changes in practice within the earth and environmental sciences over the past twenty years. These changes, catalyzed by advances in computational power and internet technologies, include a marked shift towards integrative and synthetic investigations involving collaborations of multidisciplinary researchers working with existing, heterogeneous, and potentially massive data. At the same time, the volume of data continues to grow rapidly, through new information generated by satellites and on-the-ground sensors, and the accumulation of long-term observations. Ironically, these positive trends have created significant challenges for scientists to discover and interpret relevant data, to more efficiently couple their integrative models and analyses, and to re-use and enhance/extend existing code instead of re-inventing it.

There has thus been great excitement about the potential for innovations in knowledge representation and associated reasoning methodologies to enable a new age of discovery and interoperability for integrative natural sciences. We envision these occurring through 1) vastly increased efficiency and precision in our searches for relevant data; 2) aiding in the discovery and interpretation of past results of models and analyses; 3) facilitating component re-use in scientific workflows; 4) linking together currently disparate parts of the research process, such as methods with results, or certain outcomes with related studies; and 5) uncovering latent relationships among research topics both intra- and inter-disciplinarily.

I believe the most significant current impediment to more effectively advancing the above mission is the lack of relevant cyberinfrastructure to enable community-wide uses and benefits in KR. Specifically, outside of the genomics community, ontologies and other controlled vocabularies are not well established, nor accepted and validated by the research community for most fields of earth sciences. There are many modest to significant vocabularies that have been constructed, yet closer examination reveals these are often inconsistent with one another, with incompatible axiomatic structures, displaying disciplinary quirks in representation if not outright errors, critical gaps in content, and typically having unclear or simplistic inferencing utility. For the earth sciences, a dedicated, community-based ontology construction effort is desperately needed, that allows for researcher input and vetting, working in close conjunction with KR experts, with a commitment to backward compatibility so that current investments will be useable, though not necessarily as powerful, as KR languages and reasoning engines continue to improve.

Closely related to success of the above venture, is the need for far more accessible applications and interfaces to enable the interrogation of information resources within KR systems. This would include far more visually-assisted exploration of KR-housed resources, that might typically include many levels of hierarchy and relationships that somehow need to be exposed to assist the researcher with refining their search.

Finally, the back-end systems for housing KR resources must become far more scalable, with either on the fly or far more efficient reasoning performance, as linking up massive amounts of related information could be highly useful, but would inevitably lead to a massive knowledge base that scientists might want to approach from any number of disciplinary perspectives.