Artificial Intelligence Needs Open-Access Knowledgebase Contents

Erik Sandewall
Department of Computer and Information Science
Linköping University, Linköping, Sweden
and KTH - Royal Institute of Technology, Stockholm, Sweden

A substantial knowledgebase is an important part of many A.I. applications as well as (arguably) in any system that is claimed to implement broad-range intelligence. Although this has been an accepted view in our field since very long, little progress has been made towards the establishment of large and sharable knowledgebases. Both basic research projects and applications projects have found it necessary to construct special-purpose knowledgebases for their respective needs. This is obviously a problem: it would save work and speed up progress if the construction of a broadly sharable and broadly useful knowledgebase could be a joint undertaking for the field.

In this article I wish to discuss the possibilities and the obstacles in this respect. I shall argue that the field of Knowledge Representation needs to adopt a new and very different paradigm in order for progress to be made, so that besides working as usual on logical foundations and on algorithms, we should also devote substantial efforts to the systematic preparation of knowledgebase *contents*.

The article is based on the author's experience from the ongoing development of the Common Knowledge Library (http://piex.publ.kth.se/ckl/), CKL. This is an open-access resource for modules of knowledge and facts, presently containing more than 70,000 entities with their associated attributes and values.

Factbase, Ontology and Knowledgebase

Knowledgebase efforts such as Cyc and Sumo emphasize the role of the "top-level ontology", but for the present discussion we shall take the *factbase* as the point of departure. A factbase is then a collection of information that assigns simple pieces of information to named entities, such as assigning country and region to a city, assigning local language(s) to a country or region, or assigning year and place of birth to a person.

The task of organizing a factbase can be extended in two relatively independent directions, namely, towards an ontology and towards a knowledgebase. The ontology provides a structure that helps to organize the factbase, typically including a classification system, a subsumption relation in particular for abstract concepts, and a number of formally expressed, structural restrictions.

On the other hand, the knowledgebase (in our sense of the word) contains information with a more complex structure than what can be cleanly represented in the factbase. For example, in the domain of scientific publishing, the factbase may specify the relation between a journal and the publisher of the journal, the relation to the editorial board members, and the city where the publisher is located. The knowledgebase may e.g. contain the information about the restrictions and requirements for parallell publication that different publishers have adopted, that is, the conditions under which an author may post her article on her university's website.

Knowledgebase contents are empirical in character, in the sense that it is meaningful to ask whether they are true or false, whereas the design of an ontology seems to be an analytical exercise to a large extent. In order to acquire knowledgebase contents one must therefore go out and obtain information from the real world, or rely on someone who has already done that.

The Factbase is the Base

Given that from a factbase you can proceed either to an ontology or to a knowledgebase, I propose that the continuous spectrum from empirical facts to empirical knowledge is the most important part. It is important because *the interpretation of knowledge sources at one point on that spectrum often depends on using already acquired facts and knowledge that is below it in complexity*.

Consider, for example, the seemingly elementary task of identifying the semantic contents of a university name. Compare:

- Cracow University of Technology
- Aristotle University of Thessaloniki

These two names are of course not analogous, but it takes some factual knowledge to see this. For a less simple case, consider the following university names in other languages:

- Vysoké učení technické v Brně
- Semmelweis Egyetem
- Univerzita Karlova v Praze
- Mendelova zěmědelská a lesnická univerzita v Brně
- Univerzita Konštantína Filozofa v Nitre

The semantic contents can be guessed in some cases where the English-language word is similar, but it takes factual knowledge to interpret *Semmelweis* (a scientist) and linguistic knowledge to interpret *zěmědelská* (agricultural). A combination of both is required to interpret *Praze* (Prague, in Czech and in locative case). An interesting difficulty is that the common-sense interpretation leads astray for the last line, where *Filozofa* does not indicate a philosophy-oriented university; instead, the University of Nitra has been named after Konstantin the Philosopher.

In a longer perspective we should be able to have systems that acquire knowledgebase contents by reading continuous text. Intermediate goals may then be to build systems that can identify the semantic contents of *short texts*, starting with "long names" consisting of several words, such as the names of scientific journals or the names of universities. Captions of figures and diagrams is another example of short texts that may also be addressed. It is in this context that the examples shown above are relevant.

Therefore, *in order to proceed in a systematic way towards the acquisition of contents for common, large, complex knowledgebases, it is important to first devote sufficient efforts to the factbase and to other layers of lesser complexity*. Moreover, for full coverage it is important to *combine factual and linguistic information* in these lower levels, both because many entities have non-English names, and because of the interest in using sources expressed in languages other than English.

Sources for Acquiring the Factbase

But isn't the information for such a factbase already available, and isn't it trivial to just convert it to the markup language, logic-based KR language, or any other representation system that one chooses to use? Unfortunately, the answer is no; let me briefly explain why. In the course of building the CKL I have used the following relatively large sources:

- The Dbpedia, which contains information in RDF that has been extracted from the Wikipedia
- Middle-level parts of the Sumo ontology
- The Wordnet, which contains nodes corresponding to individual "word meanings" in English
- Files from the U.S. Census in year 2000, containing information about around 25,000 "places".

In addition, several medium-sized databases and a considerable number of HTML webpages have been used.

Among these, the Dbpedia stands out as the most rewarding source. It is large and comprehensive, and its original authors have given good attention to correct nameforms and correct spelling of names (including diacritics). Time-dependent facts are often provided with a timestamp.

The Wordnet, by comparison, is less useful as a source for facts, and of course this is also not its main purpose. Its circa 20,000 nodes for proper names and adjectives (among 120,000 total) may be compared with the 80,000 entities just for persons, in the Dbpedia.

The Sumo modules use a much more expressive representation language (KIF) than the others, but it has much less information, the selection of information seems random, and it contains numerous factual errors. Therefore it is more important as a source of ontology than for its factbase.

Other sources tend to be more special-purpose, and in the case of HTML-encoded sources it is often quite messy to extract the desired information on a clean form.

There are two problems, according to this experience, if existing sources are to be used towards a universally available knowledgebase: *coverage* and *representation*. The coverage problem is that many kinds of facts are not available at all. For example, there does not seem to be any knowledge source for things and phenomena around a house, ranging from air conditioners to zippers. An ontology and factbase for this domain could however be a useful resource for home robotics projects.

Representation Problems in Dbpedia

The representation problem will be illustrated with examples from the Dbpedia, for the very reason that the Dbpedia is in many ways the best one of the sources I have been looking at. Remember, however, that since the Dbpedia has been extracted from the Wikipedia, its contents were originally written in order to be plugged into scripts for web pages. Considering them as a knowledgebase constitutes a second use that they were not originally intended for. The following examples of problems must therefore not be understood as criticism, but as indications of remaining work.

One of the largest-volume entitytypes in the Dbpedia is for cities and towns. This information is potentially useful in many applications, as different as travel planning and geographical economics. Consider the following attribute-value pairs in the Dbpedia description of the German town of Ĥerten (the notation has been changed in unimportant respects):

```
[latDeg 51]
[latMin 36]
[lonDeg 7]
[lonMin 8]
[hoehe 75]
[flaeche [unit "37.31" 2001/Decimal]]
[einwohner 64344]
[stand [unit "2006-12-31" 2001/Date]]
[plz "45699, 45701"]
[vorwahl "0 23 66, 02 09 (Westerholt),
          0 23 65 (Marl)"]
[kfz "RE"]
```

One striking feature in this example is the use of German-language words for the attributes. This includes both full words, such as e̊inwohner (inhabitants), and acronyms such as p̊lz for Postleitzahl (zip code). Descriptions of towns in other countries use other attribute names. This means that any query facility or other service that uses this information must be able to resolve such differences at runtime, and one may wonder whether it would not be better to normalize the representation once and for all.

The tradeoff between static and dynamic reconciliation of representations is more severe in those cases where a composite concept has been serialized into an attribute value

as a string. The example shows this in the zipcode attribute where the value is intended to enumerate 45699 and 45701, and in the attribute for telephone area codes (v̂orwahl) where apparently different parts of Herten have different area codes.

The representation of composite concepts in strings leads to a need for decoding at lookup time. It also leads to a lack of uniformity that may be quite difficult to interpret; retrieval routines will have to be aware of a variety of representational conventions.

There are also several other, equally ad-hoc ways of representing composite concepts. For Herten, the timestamp of the population figure is expressed in the separate attribute called ŝtand. Compare the representation for the Catalan town of Sitges:

```
[pop1900 3162]
[pop1930 6962]
  ...
[pop1986 11889]
[popLatest 24470]
[latestYear 2001]
```

The geographical location of Herten is expressed in terms of degrees and minutes on the arc, using the two attributes l̂atDeg and l̂atMin, and similarly for the longitude. Again, different conventions are used for towns in different countries.

What is interesting however is that most of these divergences need not be a problem in the situations that the dataset in question was designed for, namely, for filling fields in the scripts that are used to generate wikipedia web pages. The problems arise *when information that was organized for being used in web pages is to be reused in a knowledgebase*.

Are These Important and General Problems?

One valid question is how general are the problems that were shown above. With respect to other parts of the Dbpedia, similar problems appear elsewhere, although of course more or less severe. For example, looking at the descriptions of famous scientists, even the expressions for the areas of Nobel prizes are expressed differently for different persons, which again adds to the requirements on auxiliary information that is to be used at lookup time. The words or phrases for the scientific discipline of a scientist suffer a similar lack of standardization.

Other sources may have problems of other kinds. For example, for a knowledgebase and from a continental European perspective, Wordnet has the weakness of a strong English-language bias. This is natural from the point of view of *its* goals, but it is a handicap when used for a knowledgebase with international relevance.

The problems that were shown in these examples are more or less trivial when taken one by one. The point is, however, that *a very large number of trivial problems is a nontrivial problem*. It may be addressed by organizing a lot of manual work, or by identifying underlying principles and developing automatic or semiautomatic tools for solving them.

The Knowledge Workbench

My main theme is that presently available datasets should be viewed as raw materials for forthcoming knowledgebases: they are the valuable results of much work, they have extensive contents, but they need additional refinement, transformation, correction, and validation before they will be satisfactory resources for continued research in our field. Moreover, this work will require a combination of automatic operations and manual interventions when working with selected parts of the given resources.

A lot of work will be needed for this: work to create knowledgebase contents for domains where it is not already available, and work on processing existing contents that were designed for other purposes so that it becomes adequate also from our point of view.

In fact there is no sharp borderline between those activities. They can be unified through the concept of a *knowledge workbench*, that is, a working situation where knowledgebase contributions are produced and validated by a combination of several activities: entering information directly, scrubbing available webpages, importing from databases, and using already accumulated knowledgebase contents and software tools in order to interpret, double-check and transform recently acquired information.

Publication of Knowledgebase Modules

However, in order for this to happen, there must also be *academic recognition* for such work, and in order to ascertain the quality of the results there must be a well-defined *quality control scheme*. My proposal is that we should establish a procedure and a venue for *publication* of knowledgebase modules that is separate from, but analogous to the system for publication of research articles. This means that a person that has prepared a moderate-sized collection of entity descriptions for entities in a particular domain should be able to submit it and have it checked (by peer review and by other means) and then made publicly available with a stamp of approval.

Publication of knowledgebase modules will differ from publication of articles in some ways, however. The criteria for formal correctness according to a type system or ontology should be made precise, and should be checked computationally for each submission. Version management will be important. The choice of publication notation will also be an issue.

The emphasis on knowledgebase modules represents a change of perspective from current thinking, away from integrated, query-oriented databases and the webpages that serve as their front-ends, and towards independently authored and validated knowledgebase components.

The Common Knowledge Library is a prototype system showing how this can be done. It is not presented as a facilty for query and browsing. Instead, CKL-published modules should be seen as knowledge-containing counterparts of open-access software modules. They are intended to be downloaded by users that include them in knowledge-based systems or other software systems that they are developing, just like software modules can be imported today.

Brief Curriculum Vitae for Erik Sandewall

Education

- B.Sc., Uppsala University, 1964
- Graduate student at Stanford, 1966-1967
- Ph.D., Uppsala University, 1969

Employment

- Docent, Uppsala University, 1969-1975
- Visiting associate professor, MIT AI Lab, 1974-1975
- Professor of computer science, Linköping University, Sweden, since 1975
- Visiting senior scientist, LAAS, Toulouose, France, 1993-1994

Honors

- Member of the Royal Swedish Academy of Sciences
- Member and vice-president of the Royal Swedish Academy of Engineering Sciences
- Honorary doctor at University of Toulouse Paul Sabatier
- Fellow of the German Research Center for Artificial Intelligence (DFKI) and past member of its Scientific Board
- Past president of the Scientific Committee of the French National Institute for Research in Computer Science and Control (INRIA)
- Fellow of the AAAI and of the ECAI

Publishing

- Started the Electronic Transactions on Artificial Intelligence and was its editor-in-chief during the first years.
- Started Linköping University Electronic Press and was its director until recently
- Past Co-Editor-in-Chief of the Artificial Intelligence Journal

Research projects (completed, selection)

- Coordinator of the A.I. branch of the European Prometheus project (Information technology in automobiles)
- Director of the WITAS Project (Intelligent UAV Technology)

Research History

My major topic of interest through the years has been the Representation of Knowledge, with particular attention to nonmonotonic logic and to reasoning about actions and change. My premier contribution in my own opinion is the research monograph "Features and Fluents" that was published in 1994. It introduces a framework for defining and analyzing the range of applicability of proposed approaches to the frame problem, and uses it to obtain precise results for twelve such approaches. These are characterized with respect to both an upper bound and a lower bound, that is, both a sufficient condition for correct applicability (lower bound) and a condition such that if this condition is not satisfied, then there provably exist situations where the method does not give the correct answer. Moreover, in most cases the upper and lower bounds are equal, thereby obtaining an exact characterization of the range of applicability.

The following are a number of other contributions over the years (exact titles of articles in italics):

- A Markov-process approach to planning (JACM article, 1969)
- The first proposal for a logic-based approach to nonmonotonic reasoning (Machine Intelligence, 1972)
- Partial evaluation and its use for A.I. purposes (A.I. Journal article, 1976)
- A Workflow approach to modelling office information systems (several articles, 1979 and onwards)
- *A functional approach to nonmonotonic logic* (Computational Intelligence, 1985)
- *Nonmonotonic Inference Rules for Multiple Inheritance with Exceptions*. Proceedings of the IEEE, 1986
- *A Representation of Action Structures* (with Ralph Rnnquist). AAAI Conference, 1986
- *Combining Logic and Differential Equations for Describing Real-World Systems*. KR 1989
- *Filter Preferential Entailment for the Logic of Action in Almost Continuous Worlds*. IJCAI 1989

  (Articles in intervening years 1989-1995 are often subsumed by the monograph "Features and Fluents")
- *Towards the Validation of High-Level Action Descriptions from their Low-Level Definitions*. AICom, 1996
- *Assessment of ramification methods that use static domain constraints*. KR 1996
- *Logic-Based Modelling of Goal-Directed Behavior*. KR 1998
- *A Software Architecture for AI Systems Based on Self-Modifying Software Individuals*. International Lisp Conference, 2003
- *Real-time Dialogues with the WITAS Unmanned Aerial Vehicle* (several co-authors). German AI Conference, 2003
- *Integration of Live Video in a System for Natural Language Dialog with a Robot* (several co-authors). DIALOR Workshop on Dialogue, 2005
- *Systems: Opening up the process*. Nature, Vol. 441, xi, 2006. (About alternative approaches to peer review).
- *Reification of Action Instances in the Leonardo Calculus*. IJCAI Workshop on Nonmonotonic Reasoning, Action and Change, 2007
- *Extending the Concept of Publication: Factbases and Knowledgebases*. Learned Publishing, 2008. (About the Common Knowledge Library).