In Proceedings of the Twenty-Second National Conference on Artificial Intelligence (AAAI 2007) 2007.

Learning by Reading: A Prototype System, Performance Baseline and Lessons Learned

Ken Barker[1], Bhalchandra Agashe[1], Shaw-Yi Chaw[1], James Fan[1], Noah Friedland[3], Michael Glass[1], Jerry Hobbs[2], Eduard Hovy[2], David Israel[4], Doo Soon Kim[1], Rutu Mulkar-Mehta[2], Sourabh Patwardhan[1], Bruce Porter[1], Dan Tecuci[1], Peter Yeh[1]

[1]Dept. of Computer Sciences
The University of Texas at Austin
Austin, Texas, USA 78712-1188
{ kbarker, bsagashe, jchaw, jfan, mrglass, onue5, sourabh, porter, tecuci, pzyeh }@cs.utexas.edu

[2]Information Sciences Institute
The University of Southern California
Marina del Rey, California, USA 90292-6695
{ hobbs, hovy, rutu }@isi.edu

[3]The Friedland Group, Inc.
noah@NoahFriedland.com

[4]SRI International
Menlo Park, California, USA 94025-3493
israel@ai.sri.com

Abstract

A traditional goal of Artificial Intelligence research has been a system that can read unrestricted natural language texts on a given topic, build a model of that topic and reason over the model. Natural Language Processing advances in syntax and semantics have made it possible to extract a limited form of meaning from sentences. Knowledge Representation research has shown that it is possible to model and reason over topics in interesting areas of human knowledge. It is useful for these two communities to reunite periodically to see where we stand with respect to the common goal of text understanding.

In this paper, we describe a coordinated effort among researchers from the Natural Language and Knowledge Representation and Reasoning communities. We routed the output of existing NL software into existing KR software to extract knowledge from texts for integration with engineered knowledge bases. We tested the system on a suite of roughly 80 small English texts about the form and function of the human heart, as well as a handful of "confuser" texts from other domains. We then manually evaluated the knowledge extracted from novel texts.

Our conclusion is that the technology from these fields is mature enough to start producing unified machine reading systems. The results of our exercise provide a performance baseline for systems attempting to acquire models from text.

Learning by Reading

Learning by reading is a term that may refer to any number of tasks involving the interpretation of natural language texts. Our task is to build a formal representation of a specific, coherent topic through deep processing of concise texts focused on that topic. This is in contrast to unrestricted text understanding, which attempts to extract as much knowledge as possible from what is explicitly expressed in given texts. Unrestricted text understanding is a much more challenging form of *learning by reading* than our task. In particular, our target topic is known and the vocabulary and required depth of the formal representation are fixed. These features of the task mean that we place more emphasis on existing background knowledge and the expectations it implies and less emphasis on "reading well". Specifically, our task allows us to read multiple texts on a topic to find the knowledge required to build a model of the topic, reducing the burden of Recall performance on any given text.

Even this restricted definition of *learning by reading* subsumes at least two subtasks that are open problems in AI—natural language understanding and knowledge integration. Advances in both areas suggest that their technologies are ready to be combined to make headway in text understanding. The use of large, domain-independent, broad-coverage, corpus-based language tools and resources has made natural language processing systems very robust. The availability of large, domain-independent, broad-coverage knowledge repositories and flexible matching techniques have made knowledge integration of *unexpected inputs* much more feasible. These advances help reduce the brittleness that has plagued end-to-end text understanding systems of the past.

Furthermore, although the subtasks are undeniably difficult, combining them might simplify both. The knowledge integration task provides a knowledge base that can supply expectations and context to help disambiguate natural language, and natural language understanding might automatically produce new content to fill gaps in the knowledge. Ultimately, with the two tasks tightly coupled in a cycle, a Learning-by-Reading system might start with only general knowledge and a corpus of relevant texts and bootstrap itself to a state of domain expertise.

This paper describes a first step toward building a Learning-by-Reading system: assembling a prototype, analyzing its performance and identifying major challenges. We built the prototype system by combining off-the-shelf systems for the tasks of parsing, semantic elaboration and knowledge integration.

The system attempts to acquire a knowledge base of CONCEPT-relation-CONCEPT triples from information conveyed by the text. The triples are not isolated facts, but are integrated with the rich logical forms of the background knowledge base. That is, each CONCEPT-relation-CONCEPT triple is part of a larger, underlying semantic graph. The result for an entire focused text is a semantic model of a topic, not simply a set of relational tuples such as those harvested by corpus-based Information Extraction tools (Banko et al., 2007). In addition to guiding integration, the background knowledge provides expectations for generating *hypotheses*—information that may be relevant to the topic, but not present in the text. These hypotheses may be useful in helping disambiguate text and guiding subsequent reading.

To test the prototype we applied it to texts in the domain of human physiology, in particular, the form and function of the human heart. The texts were in unrestricted English, were obtained from a variety of sources (including the web, encyclopedias, and biologists), and were roughly at the level of Wikipedia articles. We compared the system's output with extracted facts identified by human readers, establishing a performance baseline for evaluating future systems.

Assembling a Prototype System

We built the prototype system by integrating existing components drawn from natural language understanding and knowledge-based systems. The components had never before been combined, and they were not designed with integration in mind. We connected them in a straight pipeline architecture. For integration, the components were customized in only one way, namely to use the same set of (binary) semantic relations. The relations include:

- EVENT-to-ENTITY: agent, donor, instrument, etc.
- ENTITY-to-ENTITY: has-part, location, material, etc.
- EVENT-to-EVENT: causes, defeats, enables, etc.
- EVENT-to-VALUE: rate, duration, manner, etc.
- ENTITY-to-VALUE: size, color, age, etc.

The background knowledge base is the Component Library (Barker et al., 2001), which encodes representations of about 700 general concepts, such as the events TRANSFER, COMMUNICATE and ENTER and entities PLACE, ORGANISM and CONTAINER. The philosophy behind the Component Library is to achieve broad coverage by concentrating on general concepts in the upper ontology, but to achieve depth of representation by richly axiomatizing the relatively small number of concepts and formalizing the semantics of their composition.

To help the system get started, we seeded it with ten concepts—including PUMP and MUSCLE—that are domain-general, but important to understanding heart texts. To avoid bias we also added about twenty "confuser" concepts—including MUSICAL-INSTRUMENT and SHOE—that a naïve system might identify in texts about "organs" and "pumps".

We lightly trained the NL system by identifying novel words and phrases pertaining to the heart and adding them to the parser's domain lexicon. The exercise also identified a handful of novel syntactic patterns particular to our genre of texts.

System Components and Processing

This section describes the text analysis process and illustrates features of the prototype by way of an example paragraph (sentences numbered for clarity).

1. *The heart is a pump that works together with the lungs.*
2. *The heart consists of 4 chambers.*
3. *The upper chambers are called atria, and the lower chambers are called ventricles.*
4. *The right atrium and ventricle receive blood from the body through the veins and then pump the blood to the lungs.*
5. *It pumps blood in 2 ways.*
6. *It pumps blood from the heart to the lungs to pick up oxygen.*
7. *The oxygenated blood returns to the heart.*
8. *It then pumps blood out into the circulatory system of blood vessels that carry blood through the body.*

Natural Language Processing

The NLP system (Mulkar et al., 2007a) includes a parser, a set of rules to convert parse fragments into logical form expressions (which include logical variables and a preliminary assignment of relations), and an abductive reasoner to expand the LF expressions for knowledge integration.

Parsing. CONTEX (Hermjakob, 1997; 2001) parses NL sentences into dependency trees that also contain certain surface semantic labels (typically, case relations). CONTEX is a deterministic parser that uses a decision tree of shift-reduce parsing operation rules, which are learned from a general corpus of training sentences. We trained CONTEX further on an additional few dozen sentences that contained unusual syntactic patterns from our domain.

Logical Form Generation. The LF Toolkit (Rathod and Hobbs, 2005) generates a set of shallow logical form expressions (Hobbs, 1998). It produces a logical form fragment for each lexical item in the parse tree and uses syntactic composition relations to create variables shared by the LF fragments. Certain additional representations are introduced, for example sets associated with plurals. The result is a set of LF expressions for each part-of-speech node in the parse tree. For example, for the verb 'work' in sentence 1, the LF Toolkit uses the following rule:

work-vb(e0,x1) →
e0-work; instance-of: work; agent-of: x1,e0

where the variable *x1* represents the parse tree node for 'heart', and *e0* represents the eventuality of the working event itself. For sentence 1 the intermediate LF is:

```
is(e0,x0,x1)
 heart-nn(x0); pump-nn(x1)
```

```
work-vb(e1)
 lung-nn(x3); together_with(e2,e1,x3);
 agent-of2(x3,e1)
agent-of1(x1,e1)
```

Abductive Expansion and Reformulation. The final NL step (Mini-TACITUS) implements a version of the abductive reasoner TACITUS (Hobbs et al., 1993). It co-indexes all variables appropriately and adds additional LF expressions that can be abductively derived from the knowledge just obtained. This step uses a set of (a few dozen) hand-crafted axioms. The final LF expressions for sentence 1 are:

```
[Interpt Number: 20 (Cost: 56)
 e0-is: eventuality-of is
 x0-heart: is x1-pump; instance-of heart
 x1-pump: agent-of e1-work;instance-of pump
 e1-work: instance-of work;
  together-with x3-lung ...]
```

At this stage axioms can abductively introduce new terms and use them to connect apparently unrelated LF expressions. For example, given the two axioms

Axiom1: device(x1) & fluid(x2) → pump(x1,x2)
Axiom2: device(x1) & heart(x1) → heart(x1)

the first axiom states that a pumping activity can expect some device (named *x1*) and some fluid (named *x2*). Should the LF expressions include a device such as ‘heart’ (named, for example, *x7*) and some fluid such as ‘blood’ (named, say, *x11*), these axioms will allow the abductive unification of *x1* with *x7* and *x2* with *x11*, producing the desired path connecting ‘heart’ and ‘blood’.

The NL system passes LF expressions on to the KI system along with some housekeeping information about words and parts of speech for variables, and an indication of any “definitional language” (for example, ‘heart’ being defined in terms of ‘pump’).

Knowledge Integration

Word-to-Concept Mapping. In addition to rich semantic representations of concepts, the existing knowledge base includes links from each concept to the WordNet synsets (Miller, 1990) that most closely match the semantics of the concept. For each variable in an LF expression, the KI subsystem looks its word up in WordNet and climbs the WordNet hypernym tree (isa hierarchy) to find synsets mapped to from KB concepts. These concepts are the candidate concepts for words from the sentence. The candidates are scored on distance traveled in WordNet, depth in the KB hierarchy and type of concept (generic vs. specialized vs. domain). Preference is also given to new concepts learned in the current reading session.

For example, the noun ‘pump’ has three synsets in WordNet: `pump#1` (mechanical device), `pump#2` (heart) and `pump#3` (shoe). The KB concept PUMPING-DEVICE maps directly to `pump#1` (distance 0). No KB concept maps directly to `pump#2`, but INTERNAL-ORGAN maps to `pump#2`’s ancestor, `organ#1` (distance 2). The KB concept SHOE maps to `pump#3`’s parent `shoe#1` (distance 1). The three candidates (PUMPING-DEVICE, INTERNAL-ORGAN and SHOE) are all specialized but pre-existing KB concepts. With no other context, PUMPING-DEVICE is preferred.

Concept Creation. Each word in an LF expression is mapped to its preferred KB concept as described above. For words that directly match the name of an existing concept the system generates instances of the existing concept. For “new” words, the system may create new KB concepts. For sentence 1, KI creates a new concept for LUNG as a kind of INTERNAL-ORGAN (based on the WordNet-based semantic search).

For ‘heart’, the semantic search finds seven candidate concepts: PUMPING-DEVICE, INTERNAL-ORGAN, TRAIT-VALUE, PLACE, CONCEPTUAL-ENTITY, PAPER and SOLID-SUBSTANCE. The candidate INTERNAL-ORGAN scores highest and, in the absence of context, would be selected as superclass for the new concept HEART. The NL system, however, has identified “definitional language”, with ‘heart’ being defined in terms of ‘pump’. KI’s concept creation biases superclass selection for definitional language by asserting axioms from PUMPING-DEVICE (the top candidate for the word ‘pump’) to the instance of HEART. This instance is semantically matched against instances of each of the seven candidate concepts. INTERNAL-ORGAN is discarded in favor of PUMPING-DEVICE as the superclass for the new HEART concept.

New concept creation may be deferred if the preferred superclass for the new concept is deemed too general.

Instance Unification. Definitional language is also a cue for the KI subsystem to perform a kind of co-reference resolution: the instance corresponding to the definition target word and the instance for the genus are unified.

For example, from sentence 3 ‘The upper chambers are called atria’, ‘atrium’ is the target and ‘chamber’ is the genus. The property ‘upper’ refers to ‘chamber’ and appears as an attribute on the instance of the concept CHAMBER. Unifying the ATRIUM instance and the CHAMBER instance equates ‘upper chamber’ and ‘atrium’, transferring the property ‘upper’ to the ATRIUM instance.

Semantic Role Relabeling. The NL subsystem assigns temporary semantic relations to related terms from sentences. The KI subsystem may accept these assignments or change them based on context from the background KB.

For example, for the phrase ‘blood from the heart’ the NL system may produce the LF expression:

```
x0-heart: origin-of x3-blood
```

In the knowledge base, however, *origin-of* is a relation constrained to be between an EVENT and a SPATIAL-ENTITY (not two ENTITIES). The KI subsystem recognizes the constraint violation and searches the KB for evidence of more appropriate spatial relations between two ENTITIES. In this case, it replaces *origin-of* with *encloses*, generating the KB triple:

```
<_Heart1 encloses _Blood3>
```

The KI subsystem performs the same KB search to assign relations the NL system has missed. In the LF expression for sentence 1, *together-with* is not a legal knowledge base relation. KB search finds *agent* as a suitable reassignment:

```
<_Work1 agent _Heart0>
<_Work1 agent _Lung3>
```

Constraint Assertion. When the NL system identifies sets (via expressions of cardinality), the KI subsystem can assert set constraint axioms into the knowledge base. For sentence 2 'The heart consists of four chambers', the NL system declares the cardinality of 'chamber' to be 4 and the KI system asserts a set constraint on the *has-part* relation for the concept HEART:

```
<Heart has-part (exactly 4 Chamber)>
```

Adjective Elaboration. The existing knowledge base representation for properties allows ordering of scalar values (e.g., *hot* > *cold*), specification of values relative to classes (e.g., *tall* relative to Person), unit conversion for numeric values, etc. This representation carries significant syntactic baggage. The KI subsystem accepts impoverished property representations from NL and searches the background KB for an appropriate elaboration.

In sentence 3, the NL system produces the LF:

```
x1-chamber: property upper
```

The KI system elaborates the LF to:

```
<_Chamber1 position _Position-Value2>
<_Position-Value2 value (*upper Chamber)>
```

meaning that the instance of CHAMBER has a *position* and the value of that *position* is `*upper` relative to the collection of instances of CHAMBER.

The background knowledge base also includes information about the noun roots of denominal adjectives. If no property constants (such as `*upper`) exist for an adjective *a*, the KI system checks if there is a noun root *n* for the adjective. If so, it submits *a-h* as a potential new concept (where *h* is the head modified by *a*) that is related to the preferred candidate concept for *n*.

For example, in sentence 8 the adjective 'circulatory' in 'the circulatory system' has a noun root 'circulation'. The KI system submits CIRCULATORY-SYSTEM as a potential new concept with superclass SYSTEM, and asserts that CIRCULATORY-SYSTEM is related to FLOW (the preferred candidate concept for 'circulation').

KB Matching and Hypothesis Generation. The result of all previous NL and KI system steps is a set of triples for each sentence consisting of KB concept instances and relations between them. Coreference of concept instances is maintained, so the set of triples forms a (possibly disconnected) semantic graph. As its final "integration" step for a sentence, the KI system selects potentially relevant KB concepts (those referred to in current-sentence triples, previous triples, as well as domain concepts and learned concepts). For each relevant concept it builds an instance graph by walking KB relations to a set depth. It then chooses one concept as the best semantic match to the current-sentence triple graph using flexible semantic matching (Yeh *et al.* 2005). "Committing" to this match integrates the knowledge extracted from the sentence into a relevant part of the background knowledge base, resulting in richer, more connected (more coherent) model of the topic than the original graph of triples of the sentence.

The parts of the chosen KB concept graph that do *not* match the triple set for the current sentence are *hypotheses*. For example, in sentence 6 'It pumps blood from the heart to the lungs to pick up oxygen', the most closely matching KB concept is the PUMPING action. The sentence triples match PUMPING because there is an *instrument*, a FLUID-SUBSTANCE being pumped, an *origin* and a *destination*. Triples from the instantiated graph for PUMPING *not* accounted for in the sentence include the existence of a CONTRACT action acting on the *instrument* of the PUMPING causing a FLOW of the FLUID-SUBSTANCE from inside the *instrument* through a PORTAL to a PLACE outside the *instrument*. These facts from PUMPING form hypotheses about the heart that could be scheduled for confirmation through more reading. That is, the system could search for texts that suggest that hearts contract, blood flows, hearts have portals, etc. In the prototype system, hypotheses are reported, but not investigated.

Final Output. After processing all eight sentences, the prototype system has created eight new concepts and added them to the knowledge base: ATRIUM (a subclass of CHAMBER), BLOOD (LIQUID-SUBSTANCE), HEART (PUMPING-DEVICE), LUNG (INTERNAL-ORGAN), OXYGEN (GAS-SUBSTANCE), VEIN (BODY-PART), VENTRICLE (CHAMBER) and VESSEL (BODY-PART). It has also identified one candidate concept not added to the knowledge base: CIRCULATORY-SYSTEM (a subclass of the very general concept SYSTEM) which is related (in some unknown way) to the concept FLOW.

Also asserted to the knowledge base were 48 unique axioms. Some axioms are good:

```
<Pumping
   object Blood
   destination Lung>
<Heart has-part (exactly 4 Chamber)>
<Receive
   origin Body
   path Vein
   recipient Ventricle
   object Blood>
```

Some axioms are incorrect or overly general:

```
<Oxygen object-of Learn>
<Lung encloses Fluid-Substance>
<Entity object-of Action>
```

Evaluation

The current prototype was developed as an interactive demo system intended as a proof of concept. It was not instrumented to allow for objective comparison to other systems, or even to allow its quantitative performance to be tracked over time. The next version of the prototype, currently under development, will include a flexible question answering system and problem solver (Chaw and

Porter, 2007) to query the knowledge captured from text. Maintaining a test suite of questions will allow us to provide a quantitative measure of its performance on a real-world task. It should also allow us to compare performance to related systems in question answering, textual entailment, reading comprehension, etc.

System Performance Evaluation

We did, nonetheless, evaluate the system and its components in various ways. The main evaluation of the system's output measured its extracted information against a human *gold standard*.

We gave four novel texts on the form and function of the human heart to four human readers not associated with the project. The exercise can be thought of as an *active reading comprehension test*, where readers must identify "who did what to whom?" for each sentence. Specifically, we asked the readers to represent the content of each sentence by (1) identifying the main events described in the sentence, (2) identifying the main participants of these events, (3) deciding whether the sentence introduces new concepts outside the existing ontology, (4) identifying properties of the events and participants, and (5) describing the relationships among events and participants. For example, 'the human heart pumps blood' might be expressed as:

```
Pump (a Pumping event)                 (1)

Heart                                  (2)
Human
Blood

Heart (an Internal-Organ)              (3)
Blood (a Liquid-Substance)

Heart  part-of: Human                  (4)

Pump   instrument: Heart               (5)
       object: Blood
```

The human readers were encouraged to use their common-sense knowledge to interpret the text (e.g. to resolve anaphora), but were asked to limit their representations to include only the information conveyed by the text, either implicitly or explicitly. After the human readers worked individually to represent the texts, they discussed their encodings and collectively agreed on the final representations (the *gold standard*).

We compared the prototype system's output with the gold standard using the following metrics:

$$\textit{precision (P)} = \frac{\text{correct} + \text{partial}\times 0.5}{\text{actual}}$$

$$\textit{recall (R)} = \frac{\text{correct} + \text{partial}\times 0.5}{\text{possible}}$$

$$\textit{undergeneration (U)} = \frac{\text{missing}}{\text{possible}}$$

$$\textit{overgeneration (O)} = \frac{\text{spurious}}{\text{actual}}$$

where

correct the number of triples from the system that match a triple from the gold standard

partial the number of triples from the system that almost match the gold standard (reasonable triples that differ by at most one element)

actual total triples from the final output of the system.

possible total triples in the gold standard.

missing the number of triples in the gold standard that have no counterpart in the output of the system

spurious the number of triples from the system that have no counterpart in the gold standard

Note that partial correctness was only awarded to *reasonable* triples that differ by one element. So a triple such as `(Pump instrument Country)` would not receive partial credit, even though it differs in only one element.

Table 1 shows the results of comparing system performance to the gold standard on four test texts. The first row shows the scores for the system on the task of concept creation—the task of identifying what items in sentences require new concepts and finding the correct superclass in the knowledge base. The second row shows scores on the task of connecting concepts through relations to produce triples. The third row shows total scores.

	P	***R***	***O***	***U***
Concepts	.589	.644	.314	.174
Relations	.284	.218	.520	.664
Total	***.374***	***.322***	***.460***	***.542***

Table 1: System Precision, Recall, Over- and Under-generation versus gold standard human performance

The system seems to do reasonably well with concept creation. Overgeneration is mainly due to the system's "unique-word" approach to concept creation, resulting in new concepts such as ORGAN as a subclass of INTERNAL-ORGAN because the word 'organ' doesn't match the name of an existing concept. It is an interesting result that the human readers often preferred *not* to create new concepts when the semantics of existing concepts were "close enough".

The system performs more poorly on relation assignment. The low scores are due in part to the difficulty of this problem, to be sure, but also in part to our scoring scheme. For example, consider the case where the system has (over-eagerly) created a new PUMP concept as a subclass of PUMPING-DEVICE:

```
(Pumping instrument Pumping-Device)     (Gold)
(Pumping instrument Pump)                  (1)
(Pumping agent Pumping-Device)             (2)
(Pumping agent Pump)                       (3)
```

We would assign partial correctness to (1) because of the class mismatch and partial correctness to (2) for the relation mismatch. (3) would be scored incorrect (zero) for mismatching the gold standard in two ways, though it clearly more closely matches the content of the text than some random triple.

NLP-Oriented Expansion and Evaluation

One of the more interesting lessons learned in this effort was the effect of errors in Natural Language Processing on the final knowledge base (e.g., prepositional phrase misattachments resulting ultimately in knowledge base constraint violations).

An analysis of throughput indicated that the surface-level expressions produced at the end of abductive reasoning were often not sufficiently semantic for knowledge integration. In order to improve throughput, we trained the NLP subsystem to perform better on sentences about 'hearts' and 'blood' on a very large corpus.

The result was an increase of 15% in KR triples, and doubled coverage on the number of KB concepts reflecting content from the sentences (from 50% to 100%). This constituted a striking indication of the value of widening coverage through corpus-based training.

Sensitivity of the Prototype to the Domain

To test (informally) whether the prototype system was overly biased toward hearts as pumps, we processed six texts unrelated to the heart, but having something to do with 'pumps' or 'pumping' (e.g., 'bicycle pump', 'harmonium', 'shoe'). The goals were to verify that the results were comparable to those extracted from heart texts and to verify that background knowledge was not causing the system to "hallucinate" facts about hearts in texts unrelated to hearts. The results showed that the prototype system's performance on non-heart text is equivalent to performance on our suite of test texts about the heart.

Redundancy, Recall and Convergence

The machine reading task we envision is to read potentially many, redundant texts to build a model of a topic (although the prototype system does not seek out new texts to confirm hypotheses). The system is not expected to capture all of the content of any given text: facts missed from one text may be captured from subsequent texts.

We conducted an experiment to test two hypotheses: (1) that the redundancy of information over multiple texts lessens the *Recall* burden on the system for any single text; (2) that each subsequent text on a single topic will contribute *less* knowledge to the growing model (suggesting that the system will eventually converge on a model and can stop reading).

To test these hypotheses we compared the knowledge captured from single texts in isolation to the knowledge captured from the same texts in the context of having read other texts in the same domain. In every case, the incremental contribution from a text was *less* having read other texts than in isolation, suggesting that at least some of the concepts and relations in a given text are recoverable (if missed) from other texts.

Related Work

There has been little work on integrated Learning by Reading since the 1970s, when NLP and KR&R began to diverge. Subtasks of the reading problem have been investigated in depth by the different communities. Research from word sense disambiguation (Edmonds and Kilgarriff, 2003), semantic role labeling (Carreras and Màrquez, 2004), semantic parsing (Ge and Mooney, 2005), ontology discovery (Buitelaar et al., 2004), and knowledge integration (Murray and Porter, 1989) are all relevant to the attempt to build a machine reading system. There have been research efforts investigating more integration of the different aspects of the problem. Although these often serve a specific information need in a particular domain or simplify the task in one or more ways (Forbus et al., 2007; Hahn et al., 2002; Hovy, 2006; Mulkar et al., 2007b) they indicate progress in text understanding and an eagerness in the community for producing integrated machine reading systems.

Hard Problems and Next Steps

Because of the syntactic complexity of scientific text, we cannot count on parses being entirely correct. The machine reading task we envision mitigates the problem somewhat by relying on redundancy across multiple texts. An important next step will be to extend the system to perform "targeted reading", both to seek to confirm expectation-based hypotheses from the knowledge base and to take advantage of redundancy in building a model of the topic.

A promising approach we began to investigate in the prototype is to use the semantic properties of entities and relations in an inference framework to unify propositions where possible and abductively add propositions licensed by axioms, to recover the missing linkages between the independent fragments of logical form. Our experiments indicate that using abduction to overcome shortcomings in the parser helps to "mend" inadequate parses.

Learning by Reading holds several more specific challenges: word sense disambiguation, coreference resolution and a host of additional issues in semantics proper, including an adequate (though not necessarily complete) treatment of negation, modality, numerical expressions, and other phenomena. Handling discourse structure, and in general dealing with shifting focus and emphasis that often signals metonymy or quasi-metonymy, so pervasive in technical literature, is a challenge best addressed by integrating NLP with reasoning systems.

We estimate that the "semantic fragmentation" from interpreting sentences in isolation was responsible for 25% of the errors of omission by the system. We propose to investigate (1) identifying semantic coreferences in text, including indirect reference and reference across sentences, and (2) exposing information that is only implicit in text. Both of these tasks involve elaborating an interpretation using background knowledge, knowledge that typically only matches the textual interpretation imperfectly.

Building on our previous work, our focus will be on flexible matching of knowledge structures and adaptation (Yeh *et al.*, 2005, Fan *et al.*, 2005, Fan and Porter, 2004). We will continue to pursue our hypothesis that general background knowledge (like that encoded in our existing knowledge base) is key to these tasks.

Summary

In this paper we have chronicled the teaming of NLP and KR&R research groups to build a prototype Learning-by-Reading system. We described the system's components in the context of learning models of the heart from untreated encyclopedia-level natural language text. We evaluated the prototype against human performance, establishing a baseline for future systems.

The exercise taught us several interesting lessons: well-known types of NLP errors have significant consequences in representation and reasoning; there is considerable duplication in addressing common tasks between NLP and KR&R research; corpus-based training can "widen the funnel" of facts available for knowledge integration; flexible matching of automatically generated content to engineered content is essential; research in each community can assist with tasks that prove difficult when approached in isolation; it is feasible to build a system that learns formal models of topics from unrestricted text.

The modest success of the prototype encourages us to continue the experiment and work towards a more tightly integrated system more deserving of the description "Learning by Reading". More importantly, the stimulation of working across research communities has inspired us. We hope that this lesson will persuade others to "cross the line" and move us closer to the goal of machine reading.

Acknowledgments

This research was funded by DARPA's "Learning by Reading: The Möbius Project" (NBCHD030010 TO #0007).

References

Banko, M., Cafarella, M., Soderland, S., Broadhead, M., and Etzioni, O. (2007). "Open Information Extraction from the Web." *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Hyderabad.

Barker, K., Clark, P., and Porter, B. (2001). "A Library of Generic Concepts for Composing Knowledge Bases." Proceedings of *The First International Conference on Knowledge Capture (K-CAP 2001)*. Victoria.

Buitelaar, P., Handschuh, S., and Magnini, B. 2004. "Towards Evaluation of Text-based Methods in the Semantic Web and Knowledge Discovery Life Cycle." A Workshop at *The 16th European Conference on Artificial Intelligence*.

Carreras, X. and Màrquez, L. 2004. "Introduction to the CoNLL-2004 shared task: Semantic role labeling." *Proceedings of The 8th Conference on Computational Natural Language Learning.*

Chaw, S., and Porter, B., 2007. "A Knowledge-Based Approach to Answering Novel Questions." *Proceedings of the 3rd International Workshop on Knowledge and Reasoning for Answering Questions.* Hyderabad.

Edmonds, P. and Kilgarriff, A. 2003. "Special Issue on Senseval-2". *Journal of Natural Language Engineering* 9(1).

Fan, J., Barker, K., and Porter, B. 2005. "Indirect Anaphora Resolution as Semantic Path Search." Proceedings of *The Third International Conference on Knowledge Capture*. Banff.

Fan, J. and Porter, B. 2004. "Interpreting Loosely Encoded Questions." Proceedings of *The Nineteenth National Conference on Artificial Intelligence*, 399–405.

Forbus, K., Riesbeck, C., Birnbaum, L., Livingston, K., Sharma, A., and Ureel, L. 2007. "A Prototype System that Learns by Reading Simplified Texts." *Proceedings of the 2007 AAAI Spring Symposium on Machine Reading*. Palo Alto.

Ge, R. and Mooney, R. 2005. "A statistical semantic parser that integrates syntax and semantics." *Proceedings of the 9th Conference on Computational Natural Language Learning*, 9-16.

Hahn, Udo, Romacker, M., and Schulz, S. 2002. "Creating Knowledge Repositories from Biomedical Reports: The medSynDiKATe Text Mining System." *Pacific Symposium on Biocomputing*, Lihue, Hawaii.

Hermjakob, U. 1997. *Learning Parse and Translation Decisions from Examples with Rich Context.* Ph.D. dissertation, University of Texas at Austin.

Hermjakob, U. 2001. "Parsing and Question Classification for Question Answering". *Proceedings of the Workshop on Question Answering at the Conference ACL-2001*. Toulouse.

Hobbs, J., Stickel, M., Appelt, D., and Martin, P. 1993. "Interpretation as Abduction". *Artificial Intelligence* 63(1-2), 69-142.

Hobbs, J. 1998. "The Logical Notation:Ontological Promiscuity". In *Discourse and Inference: Magnum Opus in Progress.*

Hovy, E. 2006. Learning by Reading: An Experiment in Text Analysis. Invited paper. *Proceedings of the Text, Speech, and Discourse (TSD) conference*. Brno, Czech Republic.

Miller, G. 1990. "WordNet: An On-Line Lexical Database." *International Journal of Lexicography* 3(4).

Mulkar, R., Hobbs, J., and Hovy, E. 2007a. Learning from Reading Syntactically Complex Biology Texts. *Proceedings of the 8th International Symposium on Logical Formalizations of Commonsense Reasoning*. Palo Alto.

Mulkar, R., Hobbs, J., Hovy, E., Chalupsky, H., and Lin, C.-Y. 2007b. Learning by Reading: Two Experiments. *Proceedings of 3rd international workshop on Knowledge and Reasoning for Answering Questions*.

Murray, K. and Porter, B. 1989. "Controlling search for the consequences of new information during knowledge integration." *Proceedings of The Sixth International Workshop on Machine Learning*, 290-295.

Rathod, N. and Hobbs, J. 2005. LFToolkit. *http://www.isi.edu/ ˜nrathod /wne /LFToolkit /index.html*.

Yeh, P., Porter, B., and Barker, K. 2005. "Matching Utterances to Rich Knowledge Structures to Acquire a Model of the Speaker's Goal." Proceedings of *The Third International Conference on Knowledge Capture*. Banff.