Constructing a Textual KB from a Biology TextBook

Peter Clark, Phil Harrison
Vulcan Inc
505 Fifth Ave South
Seattle, WA 98104
{peterc,philipha@vulcan.com}

Niranjan Balasubramanian, Oren Etzioni
Turing Center, Dept CS & Engineering
University of Washington
Seattle, WA 98195
{niranjan,etzioni}@cs.washington.edu

Abstract

As part of our work on building a "knowledgeable textbook" about biology, we are developing a textual question-answering (QA) system that can answer certain classes of biology questions posed by users. In support of that, we are building a "textual KB" - an assembled set of semi-structured assertions based on the book - that can be used to answer users' queries, can be improved using global consistency constraints, and can be potentially validated and corrected by domain experts. Our approach is to view the KB as systematically caching answers from a QA system, and the QA system as assembling answers from the KB, the whole process kickstarted with an initial set of textual extractions from the book text itself. Although this research is only in a preliminary stage, we summarize our progress and lessons learned to date.

1 Introduction

As part of Project Halo (Gunning et al, 2010), we are seeking to build an (iPad based) "knowledgeable textbook" about biology that users can not only browse, but also ask questions to and get reasoned or retrieved answers back. While our previous work has relied on a hand-crafted, formal knowledge base for question-answering, we have a new effort this year to add a textual QA module that will answer some classes of questions using textual retrieval and inference from the book itself. As well as running queries directly against the textbook, we are also constructing a "textual knowledge base" (TKB) of facts extracted from the book, and running queries against those also. The TKB can be thought of as a cache of certain classes of QA pairs, and offers the potential advantages of allowing global constraints to refine/rescore the textual extractions, and of allowing people to review/correct/extend the extracted knowledge in a crowdsourcing style. As a result, we hope that QA performance will be substantially improved compared with querying against the book alone. Although this research is only in a preliminary stage, we summarize our progress and lessons learned to date.

There are four characteristics of our problem that make it somewhat unusual and interesting:

- we have a specific target to capture, namely the knowledge in a specific textbook (although other texts can be used to help in that task)
- the knowledge we want is mainly about concepts (cells, ribosomes, etc.) rather than named entities
- we have a large formal knowledge-base available that covers some of the book's material
- we have a well-defined performance task for evaluation, namely answering questions from students as they read the eBook and do homework

We describe how these characteristics have impacted the design of the system we are constructing.

2 Approach

We are approaching this task by viewing the textual KB as a cache of answers to certain classes of questions, subsequently processed to ensure a degree of overall consistency. Thus tasks of KB construction and question-answering are closely interwoven:

- The KB is a cache of answers from a QA system

Proc. of the Joint Workshop on Automatic Knowledge Base Construction & Web-scale Knowledge Extraction (AKBC-WEKEX), pages 74–78,
NAACL-HLT, Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

- A QA system answers questions using information in the KB

Thus this process can be bootstrapped: QA can help build the KB, and the KB can provide the evidence for QA. We kickstart the process by initially seeding the KB with extractions from individual sentences in the book, and then use QA over those extractions to rescore and refine the knowledge ("introspective QA").

2.1 Information Extraction

Our first step is to process the textbook text and extract semi-structured representations of its content. We extract two forms of the textbook's information:

Logical Forms (LFs): A parse-based logical form (LF) representation of the book sentences using the BLUE system (Clark and Harrison, 2008), e.g., from *"Metabolism sets limits on cell siz*e" we obtain:

```
(S (SUBJ ("metabolism"))
   (V ("set"))
   (SOBJ ("limit" ("on" ("size" (MOD ("cell"))))))
```

Triples: A set of arg1-predicate-arg2 triples extracted via a chunker applied to the book sentences, using Univ. Washington's ReVerb system (Fader et al, 2011), e.g., from "*Free ribosomes are suspended in the cytosol and synthesize proteins there."* we obtain:

["ribosomes"] ["are suspended in"] ["the cytosol"]

These extractions are the raw material for the initial textual KB.

2.2 Knowledge-Base Construction and Introspective Question-Answering

As the ontology for the TKB, we are using the preexisting biology taxonomy (isa hierarchy) from the hand-build biology KB (part of the formal knowledge project). Initially, for each concept in that ontology, all the extractions "about" that concept are gathered together. An extraction is considered "about" a concept if the concept's lexical name (also provided in the hand-built KB) is the subject or object of the verb (for the LFs), or is the arg1 or arg2 of the triple (for triples). For example ["ribosomes"] ["are suspended in"] ["the cytosol"] is an extraction about ribosomes, and also about cytosol, and so would be placed at the Ribosome and Cytosol nodes in the hierarchy.

As the extraction process is noisy, a major challenge is distinguishing good and bad extractions. If we were using a Web-scale corpus, we could some function over frequency counts as a measure of reliability e.g., (Banko et al, 2007). However, given the limited redundancy in a single textbook, verbatim duplication of extractions is rare, and so instead we use textual entailment technology to infer when one extraction supports (entails) another. If an extraction has strong support from other extractions, then that increases the confidence that it is indeed correct. In other words, the system performs a kind of "introspective question-answering" to compute a confidence about each fact X in the KB in turn, by asking whether (i.e., how likely is it that) X is true, given the KB.

To look for support for fact X in the LF database, the system searches for LFs that are subsumed by X's LF. For example, "animals are made of cells" subsumes (i.e., is supported by) "animals are made of eukaryotic cells". In the simplest case this is just structure matching, but more commonly the system explores rewrites of the sentences using four synonym and paraphrase resources, namely: WordNet (Fellbaum, 1998); the DIRT paraphrase database (Lin and Pantel, 2001); the ParaPara paraphrase database (from Johns Hopkins) (Chan et al, 2011); and lexical synonyms and hypernyms from the hand-coded formal KB itself (Gunning et al, 2010). For example, the (LF of the) extraction:

> *Channel proteins help move molecules through the membrane.*

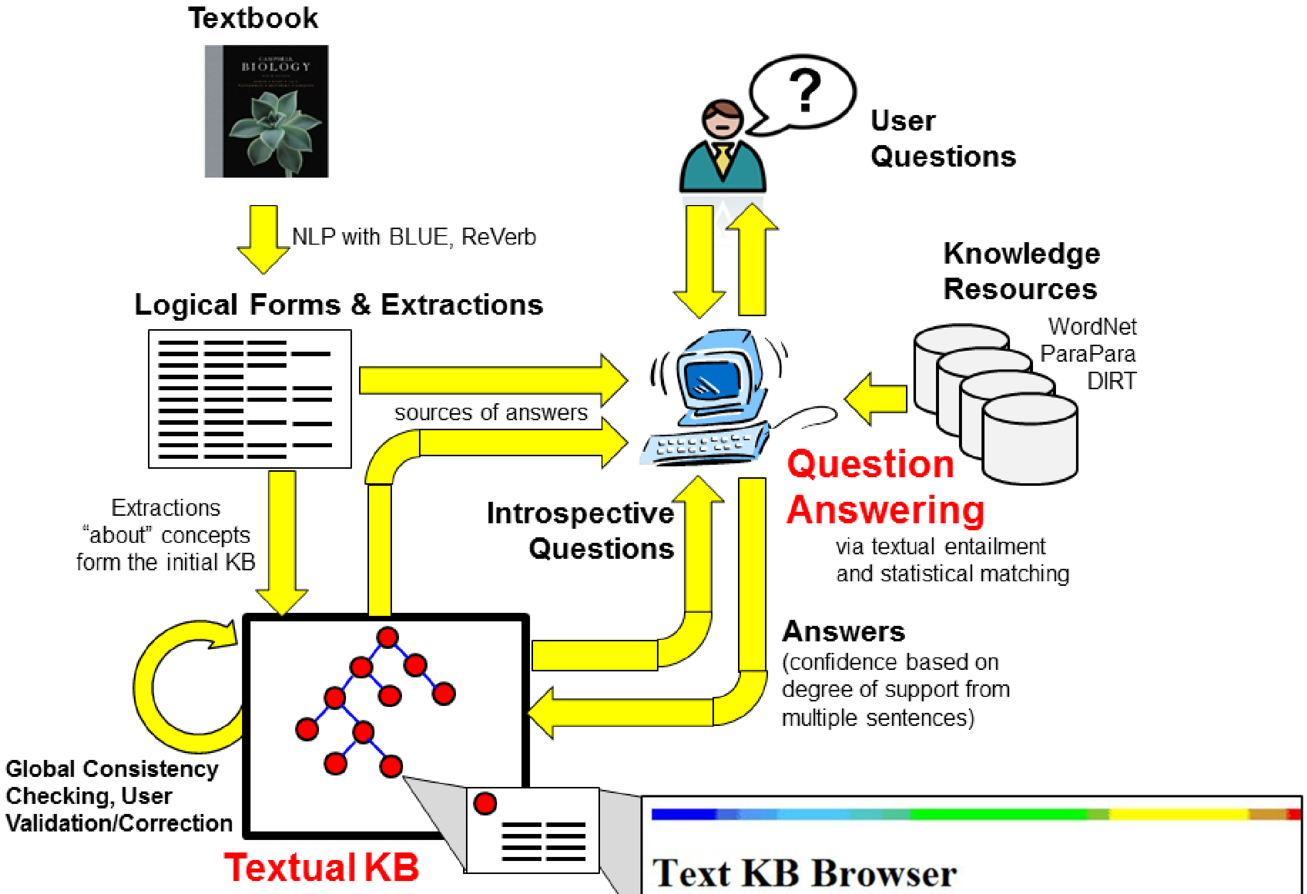
is supported (i.e., entailed) by the (LF of the) extraction:

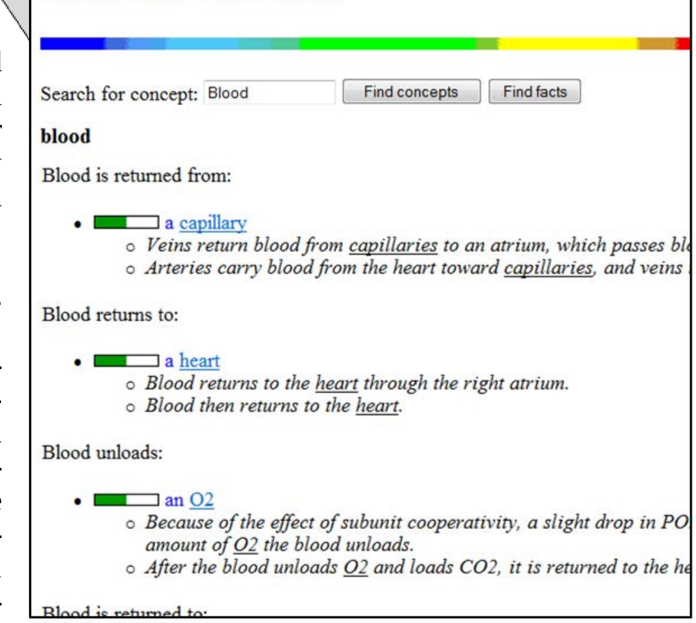
> *Channel proteins facilitate the passage of molecules across the membrane.*

using knowledge that

IF X facilitates Y **THEN** X helps Y (DIRT)
"passage" is a nominalization of "move" (WN)
"through" and "across" are synonyms (ParaPara)

To look for support for fact X in the triple database, the system searches for triples whose arguments and predicate have word overlap with the (triple representation of) the assertion X, with a (currently ad hoc) scoring function determining confidence. (Linguistic resources could help with this process also in the future).


Textbook
NLP with BLUE, ReVerb
Logical Forms & Extractions
sources of answers
Extractions "about" concepts form the initial KB
User Questions
Knowledge Resources
WordNet
ParaPara
DIRT
Question Answering
via textual entailment and statistical matching
Introspective Questions
Answers
(confidence based on degree of support from multiple sentences)
Global Consistency Checking, User Validation/Correction
Textual KB



Text KB Browser

Search for concept: Blood [Find concepts] [Find facts]

blood

Blood is returned from:

- ▬ a capillary
  - *Veins return blood from capillaries to an atrium, which passes bl*
  - *Arteries carry blood from the heart toward capillaries, and veins*

Blood returns to:

- ▬ a heart
  - *Blood returns to the heart through the right atrium.*
  - *Blood then returns to the heart.*

Blood unloads:

- ▬ an O2
  - *Because of the effect of subunit cooperativity, a slight drop in PO amount of O2 the blood unloads.*
  - *After the blood unloads O2 and loads CO2, it is returned to the he*

Blood is returned to:

Figure 1: Extractions from the textbook are used for question-answering, and a selected subset form the initial text KB. Its contents are then verified (or refuted) via introspective QA, global consistency checks, and user validation. The resulting KB then assists in future QA.

Both of these methods are noisy: There are errors in the original extractions, the synonym databases, and the paraphrase databases, not to mention oversimplifications and context-dependence in the original book sentences themselves. To assign an overall confidence to an LF-based extraction entailed (supported) by multiple sentences in the TKB, we use machine learning to build a confidence model. Each (numeric) feature in this model is a combination (max, sum, min, etc.) of the individual entailment strengths that each sentence entails the extraction. Each individual entailment strength is a weighted sum of the individual paraphrase and synonym strengths it uses. By using alternative functions and weights, we generate a large number of features. Each final class is a numeric value on a 0 (wrong) to 4 (completely correct) scale. Training data was created by six biology students who scored approximately 1000 individual extractions (expressed as question-answer pairs) on the same 0-4 scale. A page from the TKB browser is shown in Figure 1 (the bars representing confidence in each assertion).

2.3 Knowledge Refinement

We have a preliminary implementation of the first two steps. This third step (not implemented) is to refine the textual KB using two methods:

- Global coherence constraints
- User (“crowd”) verification/refinement

Our goal with global coherence constraints is to detect and remove additional extractions that are globally incoherent, even if they have apparent sentence-level support, e.g., as performed by (Carlson et al., 2010, Berant et al., 2011). Our plan here is to identify a "best" subset of the supported extractions that jointly satisfies general coherence constraints such as:

transitivity: $r(x,y) \wedge r(y,z) \rightarrow r(x,z)$
reflexivity: $r(x,y) \leftrightarrow r(y,x)$
irreflexivity: $r(x,y) \leftrightarrow \sim r(y,x)$

For example, one of the (biologically incorrect) assertions in the TKB is "Cells are made up of organisms". Although this assertion looks justified from the supporting sentences (including a bad paraphrase), it contradicts the strongly believed assertion "Organisms are made up of cells" stored elsewhere in the TKB. By checking for this global consistency, we hope to reduce such errors.

In addition, we plan to allow our biologists to review and correct the extractions in the KB in a "crowd"-sourcing-style interaction, in order to both improve the TKB and provide more training data for further use.

2.4 Performance Task

We have a clear end-goal, namely to answer students' questions as they read the eBook and do homework, and we have collected a large set of such questions from a group of biologists. Questions are answered using the QA methods described in step 2, only this time the questions are from the students rather than introspectively from the KB itself. The textual KB acts as a second source of evidence for generating answers to questions; we plan to use standard machine learning techniques to learn the appropriate weights for combining evidence from the original book extractions vs. evidence from the aggregated and refined textual assertions in the textual KB.

3. Discussion

Although preliminary, there are several interesting points of note:

1. We have been using a (pre-built) ontology of concepts, but not of relations. Thus there is a certain amount of semi-redundancy in the assertions about a given concept, for example the fact "Ribosomes make proteins" and "Ribosomes produce proteins" are both the TKB as top-level assertions, and each shows the other supports it. It is unclear whether we should embrace this semi-duplication, or move to a set of predefined semantic relationships (essentially canonicalizing the different lexical relationships that can occur).
2. Our QA approach and TKB contents are largely geared towards "factoid" questions (i.e., with a single word/phrase answer). However, our target task requires answering other kinds of questions also, including "How..." and "Why..." questions that require a short description (e.g., of a biological mechanism). This suggests that additional information is needed in the TKB, e.g., structures such as

   because(*sentence1*,*sentence2*)

   We plan to add some semantic information extractors to the system to acquire some types of relationships demanded by our question corpus, augmenting the more factoid core of the TKB.
3. We are combining two approaches to QA, namely textual inference (with logical forms), and structure matching (with ReVerb triples), but could benefit from additional approaches. Textual inference is a "high bar" to cross - it is reasonably accurate when it works, but has low recall. Conversely, structure matching has higher coverage but lower precision. Additional methods that lend extra evidence for particular answers would be beneficial.
4. We are in a somewhat unique position of having a formal KB at hand. We are using it's ontology both as a skeleton for the TKB, and to help with "isa" reasoning during word matching and textual inference. However, there are many more ways that it can be exploited, e.g., using it to help generate textual training data for the parts of the book which it does cover.
5. While there are numerous sources of error still in the TKB, two in particular stand out, namely

the lack of coreference resolution (which we currently do not handle), and treating each sentence as a stand-alone fact (ignoring its context). As an example of the latter, a reference to "the cell" or "cells" may only be referring to cells in that particular context (e.g., in a paragraph about eukaryotic cells), rather than cells in general.

6. We are also in the process of adding in addition supporting texts to the system (namely the biology part of Wikipedia) to improve the scoring/validation of textbook-derived facts.

References

Banko, M., Cafarella, M., Soderland, S., Broadhead, M., Etzioni, O. Open Information Extraction from the Web. IJCAI 2007.

Berant, J., Dagan, I., Goldberger, J. Global Learning of Focused Entailment Graphs. ACL 2011.

Carlson, A. Betteridge, J., Wang, R.C., Hruschka, E.R., Mitchel, T.M. Coupled Semi-Supervised Learning for Information Extraction. In Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM), 2010.

Chan, C., Callison-Burch, C., Van Durme, B. Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity. In Proceedings of GEometrical Models of Natural Language Semantics (GEMS-2011).

Clark, P., Harrison, P. Boeing's NLP System and the Challenges of Semantic Representation. In Proc SIGSEM Symposium on Text Processing (STEP'08), 2008.

Clark, P. Harrison, P. 2009. An inference-based approach to textual entailment. In Proc TAC 2009 (Text Analysis conference).

Fader, A., Soderland, S., Etzioni, O. Identifying Relations for Open Information Extraction. EMNLP 2011.

Fellbaum, C. "WordNet: An Electronic Lexical Database." Cambridge, MA: MIT Press, 1998.

Gunning, D., et al., Project Halo Update - Progress Toward Digital Aristotle In AI Magazine (vol 31 no 3), 2010.

Lin, D. and Pantel, P. 2001. Discovery of Inference Rules for Question Answering. *Natural Language Engineering* 7 (4) pp 343-360.