Advances in Cognitive Systems 3 (2014) 183–200 Submitted 9/2013; published 7/2014

Creating a Knowledge Base to Enable Explanation, Reasoning, and Dialog: Three Lessons

Vinay K. Chaudhri VINAY.CHAUDHRI@SRI.COM
Nikhil Dinesh NIKHIL.DINESH@GMAIL.COM
Artificial Intelligence Center, SRI International, Menlo Park, CA 94025 USA

Daniela Inclezan INCLEZD@MIAMIOH.EDU
Miami University, Oxford, OH 45056 USA

Abstract

Our work is driven by the hypothesis that, for a program to answer questions, explain the answers, and engage in a dialog just as a human does, it must have an explicit representation of knowledge. Such explicit representations naturally occur in many situations such as in designs created by engineers, software requirements created in a unified modeling language or process flow diagrams created for a manufacturing process. Automated approaches based on natural language processing have progressed on tasks such as named entity recognition, fact extraction and relation learning, but they cannot generate expressive representations with high accuracy. In this paper, we report on our effort to systematically curate a knowledge base for a substantial fraction of a biology textbook. Although this experience and the process inherently offer insights, three aspects are especially instructive for the future development of knowledge bases both by manual and by automatic methods: (1) Consider imposing a simplifying abstract structure on natural language sentences so that the surface form is closer to the target logical form to be extracted; (2) Adopt an upper ontology that is strongly motivated and influenced by natural language; (3) Develop a set of syntactic and semantic guidelines that captures how the conceptual distinctions in the ontology may be realized in natural language. Because this representation has effectively enabled reasoning, explanation and dialog, it gives a concrete target for what should be learned by automated methods.

1. Introduction

The classical approach to achieving intelligent behavior has been driven by the knowledge representation hypothesis proposed by Smith (1982): Any mechanically embodied intelligent process will comprise of structural ingredients that (1) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and (2) independent of such external semantic attribution, play a formal but causal and essential role in engendering the behavior that manifests that knowledge. In the context of this framework, an intelligent program requires a formal representation of knowledge that can be manipulated by an automated reasoner with the goal of enabling a variety of tasks, including answering questions, producing explanations and engaging in dialog.

We have recently completed a substantial knowledge engineering effort that resulted in a knowledge base called KB_Bio_101 (Chaudhri, Wessel, & Heymans, 2013; Chaudhri et al., 2013c), which represents a significant fraction of an introductory college-level biology textbook called *Campbell Biology* (Reece et al., 2011). We have used KB_Bio_101 as part of a prototype of an intelligent textbook called *Inquire*, which is designed to help students to learn better (Chaudhri et al., 2013a). *Inquire* answers questions (Chaudhri et al., 2013c), gives natural language explanations (Banik, Kow, & Chaudhri, 2013), and engages in dialog by supporting drill down.

We conducted the knowledge engineering manually for two reasons. First, none of the available automated methods could produce a representation with the level of expressiveness and accuracy that we wanted. Second, the task of creating a knowledge base is similar to drawing a figure or building a model that provides an alternative representation for information that has been stated in textual form. Such a task is inherently creative and requires substantial human input.

While designing the knowledge engineering process, we had three goals: minimize gaps in the knowledge base, achieve consensus, and catalog difficult representation issues. In our early experience of constructing a knowledge base, we had found that the single most common cause for failures in answering a question was a gap in encoding the knowledge (Friedland et al., 2004). The process needed to be systematic to prevent omissions. Although the textbook represents a consensus on a subject, because it is written in English, substantial room exists for interpretation and disagreement. Therefore, the process needed to provide a mechanism to work through ambiguities. Finally, the state of the art in formally representing textbook knowledge is still quite primitive, and we wanted to catalog the open problems in knowledge representation.

In this paper, we describe three lessons from the knowledge engineering process that we designed to addresses the above goals. These lessons are (1) reformulate sentences as universal truths so that the surface form of knowledge is closer to the knowledge to be represented; (2) use a linguistically motivated ontology into which the knowledge is extracted; and (3) use a set of syntactic and semantic guidelines that define how various conceptual distinctions are expressed in natural language. The techniques based on these three lessons were instrumental for creating KB_Bio_101, which enabled *Inquire* to answer student questions and led to learning gains as reported in a previous paper (Chaudhri et al., 2013a). Our goal here in explaining and documenting our process is to inspire the development of both manual and automated knowledge acquisition methods that can improve upon the current process.

2. Reformulating Input Sentences

A textbook is written for pedagogical purposes. Therefore, the authors adopt a style of writing that is varied, interesting, and that tells a story. This style invariably involves first introducing concepts at an abstract level, adding more details later, and, in some cases, contradicting and/or overriding the information previously introduced.

In contrast, an automated reasoning system needs to encode knowledge only once, in a succinct manner, using sentences in a formal language. Although the axioms can be arbitrarily complex, in practice, axiom patterns frequently occur (e.g., axioms that represent necessary and sufficient properties of a concept, cardinality constrains, subclass and disjointness statements) For the purposes of

Table 1. Procedure for creating Knowledge Base content from sentences.

Textbook Sentence	Universal Truth	Concept	Plan
I. A chemical signal is detected when the signaling molecule binds to a receptor protein located at the cell's surface or inside the cell.	*During signal reception, the signaling molecule binds to a receptor protein located at the cell's surface or inside the cell.*	Signal-Reception	Signal-Reception $-$ subevent $\rightarrow$ Attach Attach $-$ base $\rightarrow$ Receptor-Protein Attach $-$ object $\rightarrow$ Molecule ...
II. The binding of the signaling molecule changes the receptor protein in some way, initiating the process of transduction.	*During signal reception, the binding of the signal molecule changes the receptor protein in some way.*	Signal-Reception	Signal-Reception $-$ subevent $\rightarrow$ Bind Attach $-$ base $\rightarrow$ Receptor-Protein$_1$ Attach $-$ result $\rightarrow$ Receptor-Protein$_2$ Receptor-Protein$_1$ $-$ has-state $\rightarrow$ Receptor-Protein$_2$
	During cell signaling, the binding of the signaling molecule initiates the process of transduction.	Cell-Signaling	Cell-Signaling $-$ subevent $\rightarrow$ Signal-Reception Cell-Signaling $-$ subevent $\rightarrow$ Signal-Transduction Signal-Reception $-$ next-event $\rightarrow$ Signal-Transduction

the current discussion, we will work with one such axiom pattern known as universal truth (UT): a set of facts that are true for all instances of a concept.

To determine what should be represented from a textbook, a knowledge encoder must gather all the sentences that describe that concept. In general, a sentence will mention more than one concept. To determine which concept a sentence actually refers to, the encoder reformulates that sentence as a UT. A sentence may result in more than one UT. In our current process, the encoders work at the level of a single chapter. After the sentences in a chapter have been reformulated as UTs, we sort the UTs by concept, making available all UTs that describe a particular concept at one place. This process deals with the pedagogical style of the textbook by collecting information about a concept in one place in a similar surface syntax. We now illustrate this process with two example sentences (I and II) in Table 1.

2.1 From Sentences to Universal Truths

Syntactically, a UT is a statement of the form: (1) Every X Y (2) In X, Y (3) During X, Y. In these statements, X is a noun phrase denoting a concept and Y is a clause or verb phrase denoting information that is true about the concept. The concept (X) may not be directly mentioned in the sentence and it may be inferred from the preceding or following sentences.

The UT associated with sentence I has the form: "During X, Y", where the "X" is "signal reception." The phrase "signal reception" is not directly mentioned in the sentence, but is inferred from the phrase "a chemical signal is detected" based on the context in which the sentence appears in the textbook. A sentence may have more than one UT each of which is about a different concept.

2.2 From Universal Truths to Knowledge Representation Plans

When formalized in logic, each UT leads to an existential rule (i.e., a rule whose antecedent has one variable that is universally quantified, and whose consequent has one or more variables that are existentially quantified.) We convert each UT into a *plan*: a set of literals that would appear in the consequent of an existential rule. The plan for a UT is made by taking into account the plans for all its superclasses and dependent concepts. Thus, the knowledge that has been encoded for a superclass of a class does not need to be re-encoded for that class.

Consider the first UT in Table 1: "During signal reception, the signaling molecule binds to a receptor protein located at the cell's surface or inside the cell." A portion of the plan for this UT is shown in the fourth column and this can be understood as follows: (a) $\mathsf{Signal\text{-}Reception} - \mathrm{subevent} \rightarrow \mathsf{Attach}$: One of the steps of signal reception is an "attach" or "bind" event. (b) $\mathsf{Attach} - \mathrm{object} \rightarrow \mathsf{Molecule}$: The object (i.e., the entity that undergoes attachment) of the attach event is a molecule. (c) $\mathsf{Attach} - \mathrm{base} \rightarrow \mathsf{Receptor\text{-}Protein}$: The base (i.e., the entity that the object attaches to) is a receptor protein. We omit the remaining literals, which show the "signaling" role of the molecule and the location of the protein.

Taken together, these literals can be understood as: "one of the steps of signal reception is the attachment of a molecule to a receptor protein." The event Attach and the relations object and base are provided by the upper ontology called the Component Library (Barker, Porter, & Clark, 2001), which we discuss in more detail in the next section. The plans for a knowledge base are similar to the design specification of or a pseudo code for a program. Writing the plans first helps an encoder to think through the overall design of the representation before entering it into the knowledge base.

During this step, each sentence that could not be directly represented was tagged with the representation challenge. Doing so allowed us to identify the open problems and challenges in encoding knowledge. The most common challenge in this exercise was to first state a given piece of biological knowledge in computational terms. For example, when the textbook talks about the structure of an entity that is not explained directly in the textbook, the semantic relationships are meant need to be explicated (Chaudhri, Dinesh, & Heller, 2013). The challenges can be organized along two orthogonal dimensions: major areas of biological knowledge such as the energy transfer, process regulation, continuity and change; and major representation issues such as causality, negative information, and disjunctive knowledge.

2.3 From Plans to Knowledge Representation

The plans are entered into the knowledge base using a graphical interface (Clark et al., 2001). Figure 1 shows the concept graph for Signal-Reception; the white color denotes that it is universally quantified, whereas all other concepts are existentially quantified. The concept graph can be read as

the existential rule: "Every signal reception event has a subevent in which a molecule attaches to a receptor protein, resulting in a change in the state of the protein".

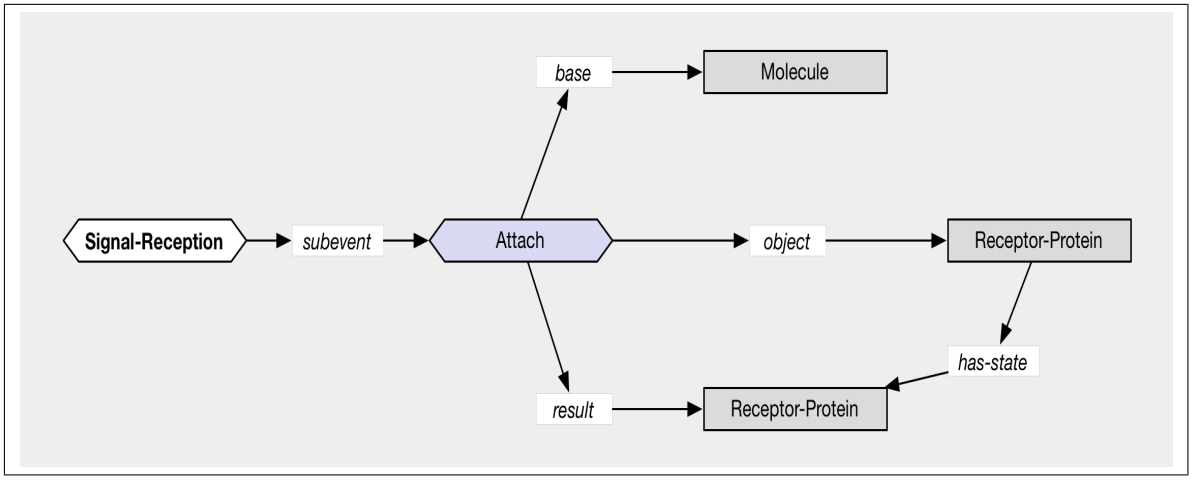

Figure 1. A partial concept graph for signal reception.

Several side benefits result from reformulating these sentences as UTs. First, the sentence form is closer to the actual logical form that will be represented in the knowledge base, making the task of creating the concept graphs much easier. Second, UTs aid in developing a consensus understanding of the content of the textbook. Finally, UTs help the encoder to think through which concepts to associate the knowledge with so that it is encoded at the most general place in the knowledge base.

2.4 Discussion

Reformulating a sentence as a UT can be viewed as a way to arrive at a surface structure of a sentence that is more closely aligned with the ultimate logical form that needs to be created. Of course, the idea of UT needs to be generalized to a broader set of axiom templates to support sufficient properties, constraints, disjointness etc. Developing such forms is the subject for future work.

In the earlier phase of our project, we did not reformulate input sentences into UTs, which led to debate and confusion about which concept to associate with a piece of knowledge. Because the knowledge was not always encoded at the most general place in the knowledge base, a lot of failures resulted (Friedland et al., 2004). The UT writing step helps an encoder to identify which variable is universally quantified and forces them to associate a piece of knowledge with the most general concept in the knowledge base. As another side-benefit, this step enables multiple encoders to reach consensus on the meaning of the sentence for which our earlier process provided no mechanism.

The significance of this approach can be understood by noting that the task of acquiring UTs is more complex than named entity recognition (McCallum & Li, 2003) and relation extraction (Carlson et al., 2010) – tasks that are the focus of current research in automated knowledge acquisition. The representation of a UT requires existential rules that, in general, can be structured as graphs

(Chaudhri et al., 2013b). An automated method must be able to identify the universal quantifier, and connect independently extracted relations into a graph; however, that cannot be done reliably. For an automated knowledge acquisition method, the availability of UTs can make the task of logical form generation substantially more tractable. Textbook sentences are so complex that unless an abstract structure such as a UT is used, the task of generating a reasonable logical form is almost impossible.

3. Linguistically Motivated Upper Ontology

Once the sentences were reformulated as UTs, we used a linguistically motivated ontology for formally representing them. We give here motivation for adopting such a strategy and the design of the ontology.

One of the most commonly used resources in natural language processing is WordNet (Miller & Fellbaum, 2007). WordNet is successful because it is linguistically motivated and encodes knowledge at the level of words, both of which ensure good coverage and facilitate understanding of what it should or should not contain. WordNet, however, is not an ontology and has several limitations regarding the support for automated reasoning (Gangemi et al., 2003).

The Component Library (CLIB) is a linguistically motivated ontology designed to support representation of knowledge for automated reasoning (Barker, Porter, & Clark, 2001). CLIB adopts four simple upper level distinctions: *entities* (things that are); *events* (things that happen); *relations* (associations between things); and *roles* (ways in which entities participate in events). We will focus on the taxonomy of physical actions where Action is a subclass of Event. Focusing on actions serves to illustrate how the library of actions is grounded in language and helps us assess coverage in a manner similar to assessing coverage for WordNet and at the same time defines the actions to support automated reasoning, explanation generation and dialog.

In the original version of CLIB (Barker, Porter, & Clark, 2001), Action has 42 direct subclasses and 147 subclasses in all. Examples of direct subclasses include Attach, Impair and Move. Other subclasses include Move-Through (which is a subclass of Move), and Break (which is a subclass of Damage, which is a subclass of Impair). To ensure generality, these subclasses were developed by consulting lexical resources, such as WordNet, the Longman Dictionary of Contemporary English (Summers, 1987) and Roget's Thesaurus (Lloyd, 1982).

We now discuss how this linguistic grounding helped us address the following two problems in our recent effort to represent knowledge from a biology textbook: (1) ensuring that we have an adequate coverage of actions that occur in the textbook, and (2) developing guidelines that inform encoders about which action from the library should be used to model a verb appearing in a sentence.

3.1 Ensuring Coverage

To check whether CLIB had coverage to support all the processes that we needed to create for the textbook, we analyzed the *verbs* appearing in the textbook. We investigated whether and how their meaning could be represented using CLIB actions and determined what new action classes should be added to CLIB when no pre-existing classes matching its meaning were found.

Table 2. Textbook verbs with a frequency higher than 400.

Freq.	Verb	Freq.	Verb	Freq.	Verb	Freq.	Verb
18,407	to be	860	to produce	629	to make	460	to increase
3,805	to have	708	to include	528	to cause	451	to grow
1,433	to call	658	to form	499	to develop	429	to become
936	to use	646	to occur	488	to do	413	to help

Campbell Biology consists of 30,346 sentences. We extracted all the verbs appearing in these sentences, which gave us a list of 2,870 verbs. The actual number of verbs is smaller, because some of the identified verbs are in fact just different forms of the same verb (e.g., *is* and *were*, two forms of the verb *to be*, were counted as different verbs). Next, we stemmed verbs on the basis of their frequency, which ranged from 1 to 18,407. The 16 verbs with a frequency higher than 400 are shown in Table 2. Some 800 verbs occurred at a frequency greater or equal to ten.

We analyzed all the verbs with a frequency greater than ten to check whether their meaning was adequately represented using some action in CLIB. We identified whether a new action class should be added, or whether we should extend the meaning of an existing class.

We identified 21 new action classes that should be added to CLIB. While adding these classes, we used the principle of correspondence (i.e., in many cases, pairs of actions go together and both should be present in the action library). For example, the initial version of CLIB contained a class called Attach referring to an *asymmetric* attachment of one entity to another, but no class existed for a *symmetric* attachment between two entities. We remedied this problem by introducing the class Bind, which is the symmetric version of Attach. We introduced the class Expel as a counterpart of Take-In, where Expel and Take-In are the subclasses of Move-Out-Of and Move-Into, respectively. Other newly introduced classes (e.g., Kill) refine the range of one of the relations in their superclasses (e.g., Kill is a subclass of Destroying a *living* entity).

The remaining proposed action classes specify the manner in which an action is performed. For instance, Fly, Run, Swim, Crawl, Hop, and Climb were added as new subclasses of Locomotion. Alternatively, manner could be described via one or more relations defined on action classes. This second option would avoid possible problems related to an increased size of the CLIB action hierarchy and the need to reorganize it. One action class whose meaning was extended is Support. Initially, this action class was defined as *"to prevent from falling,"* whereas extending its meaning by adding the expression *"or provides some other kind of structural support"* is useful for use in the domain of biology.

The discussion in this section illustrates how grounding the ontology in natural language text helped assess its coverage in relation to the knowledge that needs to be modeled, and informed us how the library should be extended.

3.2 Choosing an Action Class

When a knowledge encoder is representing a sentence that describes some process knowledge, a choice needs to be made about which action class to use. The choice must be systematic so that it is consistent across the representation of different processes across the book as well as consistent

across multiple encoders. We approached this problem by systematically analyzing how different verbs should be mapped to actions in CLIB.

For this analysis, we limited ourselves to the 800 verbs that had a frequency greater than or equal to ten. We analyzed these verbs based on their usage in the textbook, starting with the most frequent ones. For each verb, we selected a maximum of 30 sentences drawn from different parts of the textbook to ensure that we were considering representative usage. We faced two challenges in this exercise. First, because a large number of verbs have multiple meanings depending on the context in which they were used, obviously, we had to take those different meanings into consideration when choosing an appropriate CLIB action. Second, the specification of the CLIB actions contains definitions and examples related to *common sense* domains, which are not always helpful when dealing with *specialized* knowledge from the domain of biology. For example, the CLIB action Support is defined as "to prevents from falling" as illustrated by the sentence: "Tom supported the roof with a heavy beam." However, using the verb *support* in biological descriptions can also refer to a state that prevents something from changing its shape (e.g., "Intermediate filaments support cell shape").

To address these challenges, we first developed a procedure for identifying an action class by considering one fourth of the selected verbs, and then tested the procedure on the remaining verbs. We expressed this procedure as a set of guidelines for encoding verbs using CLIB actions. In this process, we realized that frequently occurring verbs, especially those with a frequency greater than 400, tended *not* to describe actually occurring action, and therefore, did not require an event to capture their meaning. This was generally not the case with lower frequency verbs. We have generated an extensive set of guidelines to handle verbs with frequency greater than ten. We illustrate the procedure with examples.

Example 1 (Choosing the Appropriate Action Class)**.** *Textbook Sentence: The groove is the part of the protein that recognizes and binds to the target molecules on bacterial walls.*

Corresponding UTs: The protein binds at the groove with the target molecules, which are situated on the bacterial walls.

Encoding: The encoder needs to choose a CLIB action class to represent the verb. CLIB contains an action class, Attach, for asymmetrical attachments. We check that the sentence describes an asymmetrical attachment by verifying that the reverse sentence *("The target molecules on the bacterial walls attach to the protein")* does not make sense. To represent this process, we use the action class Attach and assign values to the participant relations for it as follows: object = protein*;* site = Groove*; and* base = target molecules on bacterial walls. We discuss the procedure for choosing the relations in the next section.

Example 2 (Specific Guidelines for the Verb *to cross*)**.** When analyzing sentences containing the verb *to cross,* we first determined that such sentences normally translate into either *"Entity X is crossed (interbred) with entity Y"* or *"Entity X crossed entity Y."* For UTs of the first type, when the usage is in the context of a specific experiment that involves a cross, a specific action class representing that experiment should be used. In this case, conducting a cross breeding experiment is a domain-specific class to be created and maintained by the domain experts. For UTs of second

type, the relevant CLIB class is Move-Through with participant relations having the values: object = X, base = Y.

Normally, the CLIB action selected to encode a biological process is designated as its superclass. However, two exceptions exist. Sometimes the identified CLIB action describes a *subevent* of the biological process, not its superclass. Other times, a more specific action exists in the knowledge base that should be made the superclass. For example, consider the following sentence: "Most often these existing proteins are modified by **phosphorylation**, the **addition** of a phosphate group onto the protein." In this sentence, should Add be one of the subevents of Phosphorylation, or the superclass of Phosphorylation, or neither?

We address the subevent possibility first. Let us assume that we have a biological process P and we have identified a CLIB action A that could be used to model it. We use the following test to determine whether A should be a step of P or its superclass: If saying "During P, A happens" is appropriate and P is already known to have other substeps, then A should be a sub-step. If we apply these guidelines to the sentence under consideration, we notice that saying "during phosphorylation, addition happens," is appropriate, but the textbook does not describe any other subevent of phosphorylation. Accordingly, Add should not be modeled as a substep of Phosphorylation.

Next, we consider the superclass possibility. If P is a *complex* biological process and A describes just the overall outcome of P but does not capture its intricacies, then A should not be the superclass of P; this is especially valid if P has multiple steps. In this situation, a more specific biological process from the knowledge base should be selected as the superclass of P. The reasoning behind this approach is that, in such cases, the CLIB actions tends to abstract away too many of the relevant details of the biological process. The CLIB action is useful, however, for expressing the common sense definition of the process. For example, although Phosphorylation is described as an addition of a phosphate group to a protein, encoding this process as a specialization of the CLIB action Add is not a good choice because doing so would result in an overly simplified model. We prefer to make Phosphorylation a subclass of Synthesis-Reaction, which is a subclass of Chemical-Reaction and is better suited for capturing the complexity of this process.

3.3 Discussion

Several well-known upper ontologies have been used to create knowledge bases and their goals and coverage with CLIB. Other commonly used upper ontologies are: Basic Formal Ontology (BFO) (Spear, 2006) containing 36 classes in total; General Formal Ontology (GFO) (Herre, 2010) containing 79 classes; or Suggested Upper Merged Ontology (SUMO) (Niles & Pease, 2001) and the upper ontology in the Cyc system (Lenat, 1995). These upper ontologies adopt distinctions that are motivated by philosophical considerations. For example, Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) introduces distinctions such as Endurant and Perdurant, which are respectively analogous to Entity and Event in CLIB. The Cyc upper ontology has concepts such as Partially-Tangible and Partially-Intangible, which are very difficult for domain experts to understand and are not strongly tied to natural language.

CLIB was originally created to be a linguistically motivated upper ontology. The action names are grounded in language, and the semantic relationships are based on research in linguistics. As

Table 3. Definition of sample relations in CLIB with examples.

Relation	Definition	Example
agent	The entity that initiates, performs, or causes an event.	*John* **swatted** the fly.
base	The thing referenced by the event as a major or relatively fixed thing.	Vlad **attached** the sign to *the post*.
site	The specific place of some effect of an event, as opposed to the locale of the event itself.	The nurse **stabbed** the needle in *my arm* at the hospital.

we saw, the linguistic grounding of CLIB was quite effective in achieving coverage of core concepts that were needed for modeling knowledge in the biology textbook. Several concepts in CLIB capture distinctions that are not usually expressed in language. One such example is the concept of Tangible-Entity. Such concepts were problematic for natural language generation, because if such concepts appear in the output, end users fail to understand their meaning. Ideally the use of such concept names in an ontology should be minimized, and preferably, avoided. We expect CLIB to offer particular strength for natural language processing applications because of its linguistically motivated concepts and semantic relationships.

Our hope is to evolve CLIB into an inferentially valuable knowledge resource in the same way that WordNet is a lexical resource. Fully realizing this goal, however, requires sustained work. We also encourage other researchers to make their ontologies as linguistically grounded as possible.

4. Guidelines for Choosing Semantic Relations

CLIB provides two types of relations between events and entities, motivated by "case roles" in linguistics (Barker et al., 1997): participant relations (agent, base, instrument, raw-material, result and object) and spatial relations (destination, origin, path and site). CLIB provides a semantic definition of each relation, together with the common sense examples as shown in Table 3. In the examples, the event in boldface is related to the entity in italics.

After a CLIB action is selected, the next step is identifying the semantic relationships between the action class and its various participants. It is well known that semantic distinctions are not always directly expressed in language (**?**), making it difficult to apply the definitions of the relations as shown above. The following pairs of relations are especially difficult to distinguish: agent and instrument; raw-material and instrument; base and path.

If the choice between these relationships is not made consistently and correctly, it significantly interferes with the system's ability to generate good natural language sentences to support explanation generation. We consider two specific problems caused by the lack of proper usage. First, the same entity is assigned to two or more semantic relations of the same event. With such encoding, the translation into English of events is unnatural, as shown by the following automatically produced sentence "A gated channel is closed **by a stimulus with a stimulus**." This sentence results from an action Close with object = *gated channel* and agent = instrument = *stimulus*. Second, a required relation is assigned an overly general entity such as *Physical-Object* or *Tangible-Entity*. Such pro-

cess models are only partially useful in answering questions. Further, their translations into natural language are difficult for end users to understand, as in the following sentence: "A gene is moved **into an object**". This sentence resulted from an action Move-Into with object = *gene* and base = *a tangible entity*.

To address this issue, we developed a more detailed characterization of how semantic relations might be expressed in language, and how an encoder could be better supported in choosing the most appropriate relation. Such characterization involves specifying *syntactic clues* and *examples from the domain of biology*. Syntactic definitions are usually easier to follow, because they are more precise. However, the semantic relationship, base, has an irregular syntactic definition, which varies across CLIB events. Additionally, some prepositions are associated with more than one semantic relationship (e.g., *from* may indicate either a donor or an origin). For these reasons, a combined approach based on both semantic *and* syntactic definitions, as summarized in Table 4, works the best. Such an approach benefits from the advantages of both methods while diminishing their disadvantages.

For the pairs of relations that were particularly difficult to distinguish, we performed a deeper comparative analysis and provided additional guidelines, as described in Subsection 4.1. We tested these guidelines and our definitions by asking the domain experts to convert sample encodings into English sentences and then assessing whether the resulting sentences were of good quality. We consider some problematic examples from this evaluation in Subsection 4.2, together with suggestions for correcting them.

4.1 Distinguishing between Problematic Pairs of Relations

In this section, we discuss examples of relations that, as originally defined in CLIB, were too difficult to distinguish for encoders , and our approach for developing a procedure to better distinguish them.

4.1.1 Distinguishing between Relations Agent *and* Instrument

In natural language, entities denoting the agent or the instrument of an event can both be realized as the grammatical subject of a sentence, which makes distinguishing between the two difficult. Consider the following two sentences: (a) "*Birds* **eat** small seeds." (b) "*Intermediate filaments* **support** cell shape." The subjects of these sentences are mapped into the agent and instrument relations, respectively, based on the original semantic definitions of these relations, which requires the agent to be *sentient*, but not the instrument. The original definitions are: An agent is active, whereas an instrument is passive, being used by the agent if one exists. An agent is *typically* considered sentient, *if only metaphorically*, whereas an instrument need not be.

Applying these definitions and distinctions is not always straightforward because different people have different understandings of what *sentient* means. This is illustrated by the example sentence (c) "A biomembrane blocks hydrophilic compounds." Because a biomembrane is part of a living thing, whether it is sentient by itself is unclear. To solve this problem, we complemented the specifications of the two slots by adding two syntactic tests for disambiguation: (1) Transform a sentence written in the active voice into an equivalent sentence in the passive voice. The agent is the entity

Table 4. Illustrative guidelines for mapping entities into slots.

Relation	**Semantic Definition**	**Syntactic Definition**	**Biology Examples**
agent	The entity that initiates, performs, or causes an event.	(a) the grammatical subject of a sentence in active voice (b) preposition: *by* (sentence in passive voice)	*A virus* **enters** a cell. A cell **is penetrated** *by a virus*.
object	The entity that is acted on by an event; the main passive participant in the event.	(a) the grammatical object of a sentence in active voice (b) preposition: *of*	A virus **enters** *a cell*. *A cell* **is penetrated** by a virus. ... **the penetration** *of a cell* by a virus.
instrument	The entity that is used (by the agent if there is one) to perform an event.	preposition: *with* / preceded by: *using*	An animal **walks** *using its legs*.
raw-material	The entity/ material used as input for an event.	(a) the grammatical object of verbs: *to use*, *to consume*, etc. (b) preceded by: *using*	The Calvin cycle **uses** *the ATP and NADPH* to produce sugar. *Water* **is converted** to hydrogen.
result	The entity that comes into existence as a result of an event.	(a) the grammatical object of verbs: *to produce*, *to create*, etc. (b) preposition: *to* / preceded by: *producing*	Plants **produce** *their own sugars* by photosynthesis. Water **is converted** *to hydrogen*.
donor	The entity that releases the object of an event (possibly unintentionally).	preposition: *from*	Heat **is transferred** *from the warmer body* to the cooler body.
recipient	The entity that receives (takes possession of) the object of an event.	preposition: *to*	Heat **is transferred** from the warmer body *to the cooler body*.
base	An entity that the event references as something major or relatively fixed.	*Irregular – depends on the verb.*	Water **moves** *into a cell*. Water **moves** *out of a cell*. A signal molecule **attaches** *to a receptor protein*.
origin	The place where an event (typically a movement) begins.	preposition: *from*	Water **moves** *from a hypotonic solution* to a hypertonic solution.
destination	The place where an event (typically a movement) ends.	preposition: *to*	Water **moves** from a hypotonic solution *to a hypertonic solution*.

preceded by the preposition *by*, if such an entity exists. For example, by transforming the sentence (a) into an equivalent sentence in the passive voice, we obtain: "Small seeds **are eaten** *by birds*." The noun *birds* is preceded by the preposition *by*, hence it must indicate the agent. (2) If the subject of a sentence can be replaced by a phrase containing the preposition *with* or *using* when the sentence is transformed into its passive voice equivalent, then that entity is an instrument. For instance, the sentence "Cell shape **is supported** *using intermediate filaments*" sounds natural, so *the intermediate filaments* are the instrument in sentence (b). By performing these syntactic tests on sentence (c), and using the semantic definitions above, we can determine that *the biomembrane* should be the agent of the described event.

4.1.2 Distinguishing between Relations Raw-material *and* Instrument

Consider the two sentences: (a) "A planarian **detects** light *using a pair of eyespots*." (b) "The Calvin cycle **produces** sugar *using ATP and NADPH*." Here, the preposition *using*, normally associated with the instrument relation, appears in both of the sentences. However, only (a) specifies an instrument, whereas (b) specifies a raw-material.

To determine what sets the two cases apart, we analyzed several sentences that contained verbs such as *to use*, *to produce*, *to form*, and *to consume*. We determined that two guidelines can be used to capture how these distinctions are expressed in language: (1) A raw-material is an entity that is used up in an event and does not exit in the same form as it entered the process. (2) An instrument is an entity that facilitates the occurrence of the event, but the process does not consume it. This new definition clarifies why the sentence about the Calvin Cycle is an example of a raw-material: ATP and NADPH are used up by this cycle.

4.1.3 Distinguishing between Relations Base *and* Path

Consider the sentence (a) "A molecule moves through *the cell membrane*," which describes a Move-Through action. According to the original CLIB guidelines for Move-Through, *the cell membrane* should be mapped into the base relation. This conflicts with the syntactic guidelines in Table 4, which indicate that *the cell membrane* should be the path, because it is preceded by the preposition *through*. Opting for either of the two choices causes problems as we discuss below.

Let us assume that we opt for using the slot base in the sentence (a), and let us consider a new sentence (b) "A molecule moves into *the cell*." According to the CLIB guidelines for action Move-Into, *the cell* in (b) should be the base of a Move-Into event. This leads to conflicting definitions for the slot base: in the parent class Move-Through it must be the Barrier crossed; in the subclass Move-Into it must be a Container into which an object is moved.

If we use slot path in sentence (a), then we run into a different problem, as shown in sentence (c) "A molecule moves through *a pore* of the cell membrane." For representing sentence (c), no relation would exist to assign to *the pore*, given that the slot path – the most natural choice – is already assigned the value *the cell membrane*. This is a more important issue than the first option.

To remedy this problem, we decided to allow the slot base to have different definitions for different action classes, even if these action classes are connected by subclass relationships in the CLIB ontology. The new general definition of base says that it must be "a major or relatively fixed

thing that the event references," and that it cannot be associated with other slots. More specific definitions are given in relation to each action class for which this relation is relevant.

4.2 Testing the Relation Selection Guidelines

To test the guidelines described above, we asked the encoders to apply them to encode a few representative actions and then to manually convert them back into English. Such a task directly supported our goals of enabling explanation and dialog. In most cases, the guidelines were effective (i.e., when they were followed, the resulting representations led to good natural language sentences). In this section, we will discuss only those cases where the guidelines were not effective and suggest solutions for improving them.

Consider the sentence: (a) "Liquid is transported by a eukaryotic cell to cytoplasm **inside a vesicle** through a plasma membrane using an organic molecule." In sentence (a), the *vesicle* is mapped into the instrument slot. From a syntactic point of view, the preposition *inside* normally indicates association with the base slot. However, in the process of pinocytosis, the vesicle functions more like a carrier that transports the liquid. Thus, semantically, the vesicle is closer to an instrument. The instruments are indicated by the expression *using*, which is also associated with raw-material. The encoder used the preposition *inside* for the instrument because the *using* relation had already been used to capture the raw-material in this sentence. One suggestion would be to use the expression *consuming* for the raw-material and the preposition *using* for the instrument, resulting in a new sentence "Liquid is transported by a eukaryotic cell to cytoplasm using a vesicle and consuming an organic molecule."

Consider sentence (b) "A cell recognizes another cell (a target cell) **at a plasma membrane**." Here, the *plasma membrane* is assigned the relation base, while the preposition *at* is normally related to the slot site. Semantically, this means that Cell-Cell-Recognition is a function of the plasma membrane. According to the guidelines for the modeling of *Functions* (Chaudhri, Dinesh, & Heller, 2013), this information would be modeled by making the has-function slot of the plasma membrane point to Cell-Cell-Recognition. Then, the plasma membrane can be assigned the role of site in this event, as it specifies a particular place on the agent cell where the effect of recognition occurs.

Finally, consider the sentence (c) "Transferring **by an electron** from a chemical (a reducing agent) to another chemical (an electron recipient)." In this sentence, the *electron* is assigned the relation of donor, although it is preceded by the preposition *by* which is usually associated with an agent. Reduction is defined as "a reaction in which the atoms in an element accept electrons." Hence, semantically, electrons are not a donor (nor an agent), but rather the object of this transfer. To fix this case, we replace the preposition *by* with the preposition *of* as in: "Transferring **of** an electron from a chemical (a reducing agent) to another chemical (an electron recipient)."

4.3 Discussion

A different line of research whose goals converge with ours is the work on corpus annotation that is aimed at supporting natural language processing. Example projects where that approach is being pursued include PropBank (Palmer, Gildea, & Kingsbury, 2005), TimeML (Pustejovsky et al., 2005)

and SpaceML (Cristani & Cohn, 2002). The main difference between our work and the corpus annotation projects is that we are interested in synthesizing knowledge representations that may span multiple sentences, abstract away from the linguistic structure of sentences, and are directly suitable for automated reasoning.

PropBank is "a corpus of text annotated with information about basic semantic propositions." (Lopatkovà et al., 2005) The goal is to define a methodology for mapping nouns in a sentence into *arguments* of the verb in that sentence. PropBank arguments loosely correspond to relations of CLIB, but may reflect the meaning of one or more CLIB relations (e.g., Arg0 denotes both agents and experiencers). One of the resources used by annotators of PropBank texts is a database describing the arguments associated to each verb in a selected vocabulary. For example, the arguments specified for the verb *to move* are (a) Arg0: mover (b) Arg1: moved, and (c) Arg2: destination. The task that we address is more difficult than that of PropBank.

A natural question is whether the semantic relationships used in CLIB are still appropriate or whether relationships similar to the ones used in PropBank or Framenet (Baker, Fillmore, & Lowe, 1998) should be used. There is no consensus about which semantic relationships are most appropriate (Marquez et al., 2008). It is important to make a specific choice for a given application and establish guidelines for their usage.

The knowledge engineering literature frequently provides semantic definitions of relationships, but syntactic guidelines for expressing those relationships in language are not provided. One innovative aspect of our work has been the application of the guidelines of the sorts considered in the annotation projects such as PropBank to a knowledge engineering project. We developed both syntactic and semantic guidelines that helped encoders determine which semantic relationship is most appropriate for use in a process description. The linguistically motivated semantic relationships offer the strength of being general across multiple domains.

5. Summary and Conclusions

The techniques we describe in this paper were quite effective in defining the scope of what should be represented in the knowledge base. In accordance with our knowledge engineering process, we systematically examine each sentence, provide a way to map the concepts in the sentence into the concepts in the ontology, and offer a set of guidelines to identify semantic relationships conveyed by each sentence. Once a chapter has been encoded using the process presented here, for each sentence in that chapter, one can identify how the sentence is represented in the knowledge base. Any failures can then be attributed to the reasoning capability of the system or to knowledge that was not represented because of unsolved research problems in knowledge representation.

We do not expect complete automation of the process described here to be feasible in the near future. Instead, a more likely scenario is the development of semi-automated tools. Future investigation should assess whether net savings in the knowledge engineering effort result from by first conducting some automated processing followed by manual refinement.

The work reported in this paper has been driven by the assumption that an explicit representation of knowledge is critical for a system to support reasoning, explanation, and dialog. We described key aspects of creating a knowledge base from a biology textbook. Although we used specific ex-

amples from our project, three broad lessons are of interest to other projects using both manual and automated techniques for knowledge acquisition: (1) reformulating the sentences so that their abstract structure is closer to the logical form to be acquired; (2) using a linguistically motivated upper ontology; and (3) using a combination of syntactic and semantic guidelines to specify how ontological distinctions are expressed in language. We hope that the three lessons at a general level, and the specifics of the guidelines, will inspire a new breed of manual, semi-automatic, and fully automatic tools for creating knowledge representations that are well suited for reasoning, explanation and dialog.

Acknowledgements

This work has been funded by a contract from Vulcan Inc. and by an internal award from SRI International.

References

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The Berkeley FrameNet project. *Proceedings of the Seventeenth International Conference on Computational Linguistics* (pp. 86–90). Stroudsburg, PA: Association for Computational Linguistics.

Banik, E., Kow, E., & Chaudhri, V. K. (2013). User-controlled, robust natural language generation from an evolving knowledge base. *Proceedings of the Fourteenth European Workshop on Natural Language Generation* (pp. 20–29). Sofia, Bulgaria: Association for Computational Linguistics.

Barker, K., Copeck, T., Delisle, S., & Szpakowicz, S. (1997). Systematic construction of a versatile case system. *Journal of Natural Language Engineering*, *3*, 279–315.

Barker, K., Porter, B., & Clark, P. (2001). A library of generic concepts for composing knowledge bases. *Proceedings of the First International Conference on Knowledge Capture* (pp. 14–21). New York: Association for Computing Machinery.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr., E. R. H., & Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence* (pp. 1306–1313). Menlo Park, CA: AAAI Press.

Chaudhri, V. K., Cheng, B., Overholtzer, A., Roschelle, J., Spaulding, A., Clark, P., Greaves, M., & Gunning, D. (2013a). Inquire biology: A textbook that answers questions. *AI Magazine*, *34*.

Chaudhri, V. K., Dinesh, N., & Heller, C. (2013). Conceptual models of structure and function. *Proceedings of the Second International Conference on Advances in Cognitive Systems* (pp. 255–271). Baltimore, MD: Cognitive Systems Foundation.

Chaudhri, V. K., Heymans, S., Wessel, M., & Tran, S. C. (2013b). Object-Oriented knowledge bases in logic programming. *Technical Communication of International Conference in Logic Programming*.

Chaudhri, V. K., Heymans, S., Wessel, M., & Tran, S. C. (2013c). Query answering in object oriented knowledge bases in logic programming. *Proceedings of the Sixth Workshop on Answer Set Programming and Other Computing Paradigms* (pp. 81–96). Online proceedings: Computing Research Repository.

Chaudhri, V. K., Wessel, M. A., & Heymans, S. (2013). KB_Bio_101: A challenge for OWL reasoners. *Proceedings of the Second OWL Reasoner Evaluation Workshop* (pp. 114–120). Ulm, Germany: CEUR Workshop Proceedings.

Clark, P., Thompson, J., Barker, K., Porter, B., Chaudhri, V., Rodriguez, A., Thomere, J., Mishra, S., Gil, Y., Hayes, P., & Reichherzer, T. (2001). Knowledge entry as the graphical assembly of components. *Proceedings of the First International Conference on Knowledge Capture* (pp. 22–29). New York: Association for Computing Machinery.

Cristani, M., & Cohn, A. G. (2002). SpaceML: A mark-up language for spatial knowledge. *Journal of Visual Languages & Computing*, *13*, 97–116.

Friedland, N. S., Allen, P. G., Witbrock, M., Matthews, G., Salay, N., Miraglia, P., Angele, J., Staab, S., Israel, D., Chaudhri, V., Porter, B., Barker, K., & Clark, P. (2004). Towards a quantitative, platform-independent analysis of knowledge systems. *Proceedings of the Ninth International Conference on the Principles of Knowledge Representation and Reasoning* (pp. 507–515). Whistler, Canada: Springer-Verlag.

Gangemi, A., Guarino, N., Masolo, C., & Oltramari, A. (2003). Sweetening WordNet with DOLCE. *AI Magazine*, *24*, 13–24.

Herre, H. (2010). General Formal Ontology (GFO): A foundational ontology for conceptual modelling. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and applications of ontology: Computer applications*, 297–345. Springer.

Lenat, D. B. (1995). CYC: A large scale investment in knowledge infrastructure. *Communications of the ACM*, *38*, 33–38.

Lloyd, S. M. (Ed.). (1982). *Roget's thesaurus*. London: Longman.

Lopatkovà, M., Bojar, O., Semeckỳ, J., Benesovà, V., & Zabokrtstỳ, Z. (2005). Valency lexicon of Czech verbs VALLEX: Recent experiments with frame disambiguation. *Proceedings of the Eighth International Conference on Text, Speech and Dialogue* (pp. 99–106). Czech Republic: Springer.

Marquez, L., Carreras, X., Litkowski, K. C., & Stenvenson, S. (2008). Semantic role labeling: An introduction to the special issue. *Computational Linguistics*, *34*, 145–159.

McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. *Proceedings of the Seventh Conference on Natural Language Learning* (pp. 188–191). Stroudsburg, PA: Association for Computational Linguistics.

Miller, G. A., & Fellbaum, C. (2007). WordNet then and now. *Language Resources and Evaluation*, *41*, 209–214.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. *Proceedings of the International Conference on Formal Ontology in Information Systems* (pp. 2–9). New York: Association for Computing Machinery.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, *31*, 71–106.

Pustejovsky, J., Ingria, B., Sauri, R., Castano, J., Littman, J., Gaizauskas, R., Setzer, A., Katz, G., & Mani, I. (2005). The specification language TimeML. In I. Mani, J. Pustejovsky, & R. Gaizauskas (Eds.), *The language of time: A reader*, 545–557. Oxford, UK: Oxford University Press.

Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., & Jackson, R. B. (2011). *Campbell Biology*. Boston: Benjamin Cummings imprint of Pearson.

Smith, B. C. (1982). *Reflection and semantics in a procedural language*. Doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Spear, A. D. (2006). Ontology for the twenty first century: An introduction with recommendations. `http://www.ifomis.org/bfo/documents/manual.pdf`.

Summers, D. (Ed.). (1987). *Longman dictionary of contemporary English*. London: Longman.