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

Vinay K. Chaudhri, Nikhil Dinesh, and Daniela Inclezan

Artificial Intelligence Center,
SRI International, Menlo Park, CA, 94025

Abstract. Our work is driven by the hypothesis that for a program to answer questions, explain the answers, and engage in a dialog just like a human does, it must have an explicit representation of knowledge. Such explicit representations occur naturally in many situations such as engineering designs created by engineers, a software requirement created in unified modeling language or a process flow diagram 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. Use of automated methods can be problematic in situations where the conceptual distinctions used by humans for reasoning are not directly expressed in natural language or when the representation must be used to drive a high fidelity simulation.
In this paper, we report on our effort to systematically curate a knowledge base for substantial fraction of text in a biology textbook [26]. While this experience and the process is interesting on its own, three aspects can be especially instructive for future development of knowledge bases by both manual and 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 guidelines that captures how the conceptual distinctions in the ontology may be realized in natural language. Since the representation created by this process has been quite effective for answering questions and producing explanations, it gives a concrete target for what information should be extracted by the automated methods.

Keywords: knowledge representation, ontologies, automated reasoning, conceptual models, knowledge acquisition from text

1 Introduction

Classical approach to achieving intelligent behavior has been driven by the knowledge representation hypothesis proposed by Smith [27]: Any mechanically embodied intelligent process will be comprised of structural ingredients that (a) we as external observers naturally take to represent a propositional account of the knowledge that the overall process exhibits, and (b) 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 that

it will enable a variety of tasks including answering questions, producing explanations and engaging in a dialog.

There are some domains such as engineering, manufacturing, and finance where structured representations are routinely created and are a part and parcel of a routine workflow. Automated methods based on natural language processing (NLP) techniques are quite effective at creating some limited forms of structured representations such as named entity extraction [21] and relation extraction [7].

We have recently completed a substantial knowledge engineering effort that has resulted in a knowledge base called `KB_Bio_101` that represents a significant fraction of an introductory college-level biology textbook [11, 10]. We have used `KB_Bio_101` as part of a prototype of intelligent textbook called Inquire that is designed to help students in learning better [8]. Inquire answers questions [10], gives explanations and engages in dialog through natural language generation [1].

In this paper, we describe three specific aspects of the knowledge engineering process and discuss the lessons that can be drawn from this effort which can inspire the development of a new breed of manual as well as automated knowledge acquisition methods. These lessons are: (1) re-formulating sentences as universal truths so that the surface form of knowledge is closer to the knowledge to be extracted (2) using a linguistically motivated ontology into which the knowledge is extracted (3) using a set of guidelines that define how various conceptual distinctions are expressed in natural language. These three techniques were instrumental in creating `KB_Bio_101` that enabled Inquire to answer students questions and led to learning gains as have been reported in a previous paper [8]. We have organized the paper by first discussing the techniques that we used in creating the knowledge representation followed by a discussion on how these can be instructive for future manual, automated as well as semi-automated knowledge acquisition methods.

2 Reformulating Input Sentences

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

In contrast, an automated reasoning system needs to encode knowledge only once, and in a succinct manner, using sentences in a formal language. While the axioms can be arbitrarily complex, in practice, there are frequently occurring axiom patterns, for example, axioms to represent necessary and sufficient properties of a concept, cardinality constrains, subclass and disjointness statements, etc. For the purpose of the current discussion, we will work with one such axiom pattern known as universal truth: 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 universal truth. A sentence may result in

more than one universal truth. In our current process, the encoders work at the level of a single chapter. Once the sentences in a chapter have been reformulated as universal truths, they can be sorted on the concept so that we now have available all the sentences that describe a particular concept which can then be used for representation. This process deals with the pedagogical style of the textbook by collecting information about a concept in one place in a similar surface syntax.

Let us now illustrate this process by taking two example sentences (numbered I and II) in Table 1.

Textbook Sentence	Universal Truth	Concept	Plan
I. A chemical signal is detected when the signaling molecule binds to a receptor protein located at the cells surface or inside the cell.	During signal reception, the signaling molecule binds to a receptor protein located at the cells surface or inside the cell.	Signal-Reception	Signal-Reception – subevent → Attach Attach – base → Receptor-Protein Attach – object → 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 → Bind Attach – base → Receptor-Protein$_1$ Attach – result → Receptor-Protein$_2$ Receptor-Protein$_1$ – has-state → Receptor-Protein$_2$
	During cell signaling, the binding of the signaling molecule inititates the process of transduction.	Cell-Signaling	Cell-Signaling – subevent → Signal-Reception Cell-Signaling – subevent → Signal-Transduction Signal-Reception – next-event → Signal-Transduction

Table 1: Procedure for creating KB content from sentences

2.1 From Sentences to Universal Truths

Syntactically, a universal truth (or a UT) is a statement of the form: (a) Every X Y (b) In X, Y (c) 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 might be inferred from the context and the teacher's understanding of biology.

The universal truth associated with sentence I has the form – "During X, Y", where the concept "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.

2.2 From Universal Truths to Knowledge Representation Plans

When formalized in logic, each universal truth leads to an existential rule, ie, a rule whose antecedent has one variable that is universally quantified, and whose consequent has one or more variables which are existentially quantified. Each universal truth is converted to a *plan*: which is a set of literals that would appear in the consequent of the existential rule suggested above. The plan for a universal truth is made by taking into account the plans for all its superclasses and dependent concepts. For example, the plan for Cell-Signaling would take into account the plan for Signal-Reception, which is a step of Cell-Signaling.

Consider the first universal truth 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 universal truth is shown in the fourth column and this can be understood as follows:

- Signal-Reception $-$ subevent $\rightarrow$ Attach – One of the steps of signal reception is an "attach" or "bind" event.
- Attach $-$ object $\rightarrow$ Molecule – The object (ie, the entity that undergoes attachment) of the attach event is a molecule.
- Attach $-$ base $\rightarrow$ Receptor-Protein – The base (ie, the entity that the object attached 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 (CLIB) which we will discuss in more detail in the next section.

The plans for a knowledge base are similar to design specification 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.

2.3 From Plans to Knowledge Representation

The plans are entered into the KB using a graphical interface called *concept maps* [12]. Figure 1 shows the concept map for Signal-Reception; the white color denotes that it is universally quantified, while all other concepts are existentially quantified. The concept map can be read as the following 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".

There are several side-benefits of reformulating these sentences as universal truths: (1) 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 (2) universal truths aid in developing a consensus understanding of the content of the textbook (3) They help the encoder in thinking through which concepts should the knowledge be associated with.

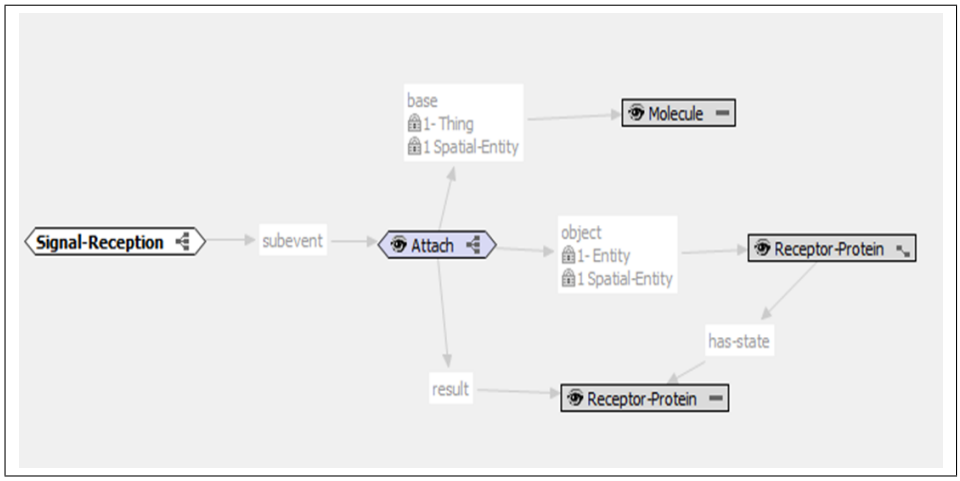

Fig. 1: Concept Map for Signal Reception

3 Linguistically Motivated Upper Ontology

Wordnet is by far the most commonly used resource in natural language processing for reasoning about entailments [22]. One of the reasons for the success of Wordnet is that it is linguistically motivated and it encodes knowledge at the level of words. This ensures good coverage and makes it easy for people to understand what it should or should not contain. Wordnet is, however, not an ontology and has several limitations when it comes to supporting automated reasoning [16].

Component Library (or CLIB) is a linguistically motivated ontology designed to support representation of knowledge for automated reasoning [3]. 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.

For the purpose of this discussion, we will focus on the taxonomy of physical actions where action is a subclass of Event. The reason for focusing on actions is 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 yet, defines the actions to support automated reasoning, explanation generation and dialog.

In the original version of CLIB [3], the Action has 42 direct subclasses and a total of 147 subclasses in all. Examples of direct subclasses include Attach, Impair, Move, and Store. 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. These subclasses were developed by consulting lexical resources, such as Wordnet [22], Longman Dictionary of Contemporary English [30] and Roget's thesaurus [20].

We will now discuss how this linguistic grounding of the ontology helped us address the following two problems in our recent effort to represent knowledge from a biology textbook: (a) ensuring that we have an adequate coverage of actions that occur in the

textbook (b) developing guidelines that inform an encoder 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 adequate coverage to support all the process representations that we will need 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 was found.

The main body of the biology textbook *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, as 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 based on their frequency, which ranged from 1 to 18,407. The sixteen verbs with a frequency higher than 400 can be seen in Table 2. There were 800 verbs with a frequency greater or equal to ten.

Verb	**Frequency**	**Verb**	**Frequency**	**Verb**	**Frequency**	**Verb**	**Frequency**
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

Table 2: Textbook Verbs with a Frequency Higher than 400

We analyzed all the verbs with frequency greater than 10 to check whether their meaning was adequately represented using some action in CLIB. As a result of this exercise, we identified whether a new action class should be added or we should extend the meaning of an existing action class.

We identified 21 new action classes that should be added to CLIB. While adding these classes, we used the principle of correspondence, ie, 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 there was no class for a *symmetric* attachment between two entities. We remedied this problem by introducing the class Bind, which corresponds to 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 re-organize it.

Finally, one example of an existing action class whose meaning should be extended is Support. Initially, this action class was defined as *"to prevent from falling,"* whereas

for use in the domain of biology it is useful to extend its meaning by adding the expression *"or provides some other kind of structural support."*

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 on which action class to use. This choice needs to 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 the purpose of 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 that contained it drawn from different parts of the textbook to ensure that we were considering representative usage. Two challenges we faced in this exercise are as follows.

1. A large number of verbs have (obviously) multiple meanings, depending on the context in which they were used. So, we must deal with different senses when choosing an appropriate CLIB action.
2. The specification of 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 instance, the CLIB action Support is defined as "to put an object in a state that prevents it from falling;" the use of this CLIB event is illustrated by the sentence:

   (1) Tom supported the roof with a heavy beam.

   However, the use of the verb *support* in biological descriptions can also refer to a state that prevents something from changing its shape:

   (2) Intermediate filaments support cell shape.

To address the above 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 an actual action taking place and therefore did not require an event to capture their meaning. This was generally not the case with lower frequency verbs. We have extensive set of guidelines to handle verbs with frequency greater than 10. For the present discussion, we illustrate the procedure by considering several examples.

Example 1. Textbook Sentence: The groove is the part of the protein that recognizes and binds to the target molecules on bacterial walls.

Corresponding Universal Truth(s): 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 *binds*. 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 will 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 will discuss the procedure for choosing the relations in the next section.

Example 2 (Guidelines for the Verb to cross*).* When analyzing sentences containing the verb *to cross*, we first determined that such sentences normally translate into UTs of one of the following two types:

(a) *Entity X is crossed (interbred) with entity Y.*
(b) *Entity X crossed entity Y.*

For UTs of type (a), whether the usage is in the context of an experiment in which an action class corresponding to 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 type (b), the relevant CLIB class is Move-Through with participant relations having the values: object = X, base = Y.

We have developed systematic guidelines to help the encoders in identifying a suitable action class from CLIB. Normally, the CLIB action selected to encode a biological process is designated as its superclass. However, there are two exceptions: sometimes the identified CLIB action describes a *subevent* of the biological process, not its superclass; other times, there is a more specific action in the KB that should be made the superclass. We illustrate this using examples.

(3) Most often these existing proteins are modified by **phosphorylation**, the **addition** of a phosphate group onto the protein.

In the above 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 it is appropriate to say "During P, A happens", and P is already known to have other substeps of P, then A should be a sub-step. If we apply these guidelines to (3), we notice that it is appropriate to say that "during phosphorylation, addition happens," but the textbook does not describe any other subevent of phosphorylation. So, 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 KB should be selected as the superclass of P. The reason 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, though, in expressing the common sense definition of the process. For instance, although Phosphorylation is described as an addition of a phosphate group to a protein in (3), encoding this process as a specialization of the CLIB action Add is not a good choice as it 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.

The discussion above illustrates the kind of procedures we needed to develop to identify suitable actions classes that should be used when modeling a process verb in a textbook sentence.

4 Guidelines for Choosing Semantic Relations

CLIB provides two types of relations between events and entities, motivated by "case roles" in linguistics [c.f. 2] :

- Participant relations – agent, base, instrument, raw-material, result, object
- Spatial relations – destination, origin, path, site.

CLIB provides a semantic definition of each relation, together with common sense examples as shown in Table 3. In the examples, the event in boldface is related to the entity in italics.

Relation	Definition	Example
agent	The entity that initiates, performs, or causes an event.	*John* **swatted** the fly
base	Event references something 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

Table 3: Definition of relations in CLIB with examples

After a CLIB action is selected for modeling some biological process described by a sentence, the next step is to identify 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 [19] 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. To further make this point, we consider two specific problems caused by lack of proper usage.

1. 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:

   (4) A gated channel is closed **by a stimulus with a stimulus**.

   The above sentence results from an action Close with object = *gated channel* and agent = instrument = *stimulus*.
2. A required relation is assigned an overly general entity such as *Physical-Object* or *Tangible-Entity*. Such process models are only partially useful in answering questions. Furthermore, their translations into natural language are difficult for end-users to understand.

   (5) A gene is moved **into an object**.

   The above 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 the 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, as they are more precise. There is however one semantic relationship, base, that has an irregular syntactic definition, which varies across CLIB events. Additionally, there are some prepositions that 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 created into English sentences and then assessing whether the resulting sentences were of good quality. We consider a few representative examples of 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 were too difficult to distinguish for encoders as originally defined in CLIB, and our approach for developing a procedure to better distinguish them.

Distinguishing between 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 it difficult to distinguish between the two:

(6) *Birds* **eat** small seeds.

(7) *Intermediate filaments* **support** cell shape.

The subjects of sentences (6) and (7) 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 the instrument need not be sentient:

- An agent is active, while an instrument is passive, being used by the agent if there is one.
- An agent is *typically* considered sentient, *if only metaphorically*, while 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 following example sentence:

(8) A biomembrane blocks hydrophilic compounds.

A biomembrane is part of a living thing, so it is not clear whether by itself, it is sentient or not. To solve this problem, we complemented the specifications of the two slots by adding some syntactic tests for disambiguation:

- Transform a sentence written in the active voice into an equivalent sentence in the passive voice. The agent is the entity preceded by the preposition *by*, if such an entity exists. (e.g., By transforming (6) 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.)
- 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. (e.g., The sentence "Cell shape **is supported** *using intermediate filaments*" sounds natural, so *the intermediate filaments* are the instrument in sentence (7).)

By performing these syntactic tests on sentence (8), and using the semantic definitions above, we can determine that *the biomembrane* should be the agent of the described event.

Distinguishing between raw-material and instrument. Consider the following sentences:

(9) A planarian **detects** light *using a pair of eyespots*.

(10) 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 (9) specifies an instrument; (10) specifies a raw-material.

To determine what sets the two cases apart, we analyzed several sentences which contained verbs such as *to use*, *to produce*, *to form*, *to consume*, etc. We determined that the following distinctions capture how these two relations are expressed in language:

- A raw-material is an entity that is used up in an event and does not come out of it the same way it entered the process.
- An instrument is an entity that facilitates the occurrence of the event, but it is not consumed by the process.

This new definition clarifies why (10) is an example of a raw-material: ATP and NADPH are used up by the Calvin cycle.

Distinguishing between base and path. Consider the sentence:

(11) 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*. However, opting for either of the two relations seems to cause problems as we discuss below.

Let us assume that we opt for using the slot base in (11), and let us consider the sentence:

(12) A molecule moves into *the cell*.

According to the CLIB guidelines for action Move-Into, *the cell* in (12) 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 that is crossed; in the subclass Move-Into it must be a Container into which an object is moved.

If we opt for using the slot path in (11), then we run into a different problem. In the sentence:

(13) A molecule moves through *a pore* of the cell membrane.

there would be no relation to assign to *the pore*, given that the slot path—the most natural choice—is already assigned the value *the cell membrane*. This is an even bigger 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 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 Our Definitions and Guidelines

To test the guidelines that we have described above, we asked the encoders to apply them to encode a few representative actions, and then manually convert them into English. Such a task is in direct support of our goals to enable explanation and dialog.

In most cases the guidelines were effective, ie, 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.

(14) Liquid is transported by a eukaryotic cell to cytoplasm **inside a vesicle** through a plasma membrane using an organic molecule. (Pinocytosis)

In (14), *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 it is closer to an instrument. Note that instruments are indicated by the expression *using*, which is also associated with raw-material. We believe that the encoder used the preposition *inside* for the instrument because the *using* relationship 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 through a plasma membrane **consuming** an organic molecule.

Next, consider the following sentence:

(15) An image is produced **using a radioactive tracer by a PET scanner**.

In (15), the *radioactive tracer* is assigned to slot agent and the *PET scanner* to the slot instrument, but the prepositions associated with the two expressions indicate a reversed assignment to slots. What happens in reality is that the image is produced by the PET device based on the computer analysis of concentrations of the tracer. Therefore, both syntactically and semantically the *tracer* should be the instrument and the *PET scanner* should be the agent.

(16) A cell recognizes another cell (a target cell) **at a plasma membrane**.

In (16), the *plasma membrane* is assigned the role base, while the preposition *at* is normally related to the slot site. Semantically, what this means is that Cell-Cell-Recognition is a function of the plasma membrane. According to the guidelines for modeling of *Functions* [9], 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.

(17) Transferring **by an electron** from a chemical (a reducing agent) to another chemical (an electron recipient). (Reduction)

In (17), the *electron* is assigned the role of donor, although it is preceded by the preposition *by* usually associated to 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).

(18) A cell receives a signal **at a receptor protein carried by a chemical**.

In (18), the *receptor protein* is assigned to slot instrument, and the *chemical* to slot object. Syntactically, the preposition *at* is used to denote the site. If we look at the definition of this process, we see that it uses a different verb than *receives*: "The target cell's *detection* of a signaling molecule coming from outside the cell." Moreover, in the encoding of this process, the chemical *plays the role* of a signal. Hence, this sentence could be reformulated as

> A chemical entity playing the role of a signal is detected by a cell using a receptor protein.

As a result, the following assignment of values to slots would be appropriate, according to the information in Table 4: object = *chemical* with plays = *signal*, base = *cell*, instrument = *receptor protein*.

5 Discussion and Lessons Learned

Let us now step back and draw some higher level conclusions from the techniques we have presented here.

Reformulating a sentence as a UT can be more generally viewed as a way to arrive at a surface structure of a sentence which 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. A closely related notion was first introduced under the name of abstract syntax trees (ASTs) [15]. UTs can be viewed as a specific instance of an AST. The use of ASTs is more broadly applicable to manual knowledge curation efforts in which the acquisition process starts from text, and an AST generation provides a graceful migration from the informal textual knowledge to a more formal logical form. In the context of automated knowledge acquisition using natural language processing methods, availability of ASTs can make the task of logical form generation substantially more tractable. The sentences in the textbook are so complex that unless one uses some form of AST, the task of getting a reasonable logical form is almost impossible. Therefore, the use of ASTs as a technique to add knowledge capture is the first major lesson or take away from the proces described here.

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

in linguistics. As 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. Even though CLIB defines semantic relationships and a few key axioms for each of the action in the library, it is far from clear how to argue the completeness of those axioms. There are several concepts in CLIB that capture distinctions that are not usually expressed in language. One such example is the concept of Tangible-Entity. As we saw during the discussion, such concepts were problematic for natural language generation, because if such concepts appear in the output, the end-users will fail to naturally understand their meaning. Ideally speaking, the usage of such concept names in an ontology should be minimized, and preferably, avoided. We expect CLIB to have special strength for natural language processing application because of its linguistically motivated concepts and semantic relationships. While we cannot claim that CLIB has yet proven its value in being an inferentially valuable knowledge resource in the same way that Wordnet is a lexical resource, continuing to develop CLIB in that direction is still a sensible direction for future work. Accordingly, we encourage and advocate other researchers to make their ontologies as linguistically grounded as possible.

Use of a combination of syntactic and semantic guidelines was essential in ensuring a systematic encoding of knowledge. We developed guidelines that helped encoders determine which semantic relationship is most appropriate for use in a process description. The linguistically motivated semantic relationships have the strength of being general across multiple domains. But, as the complexity of the guidelines indicates, they can also be difficult for humans to use and apply in a consistent manner. We hope that developing the guidelines that we presented in this paper will provide a foundation for automated and semi-automated tools that could either acquire such relationships from text automatically, or provide much better support to encoders as they make their choices. The basic idea of using a combination of syntactic and semantic guidelines is quite general and can be adopted by a broad range of applications.

6 Related Work

Several well-known upper ontologies exist today that have been used to create knowledge bases and overlap in their goals and coverage with CLIB. One of them is DOLCE [6], which is a higher-level ontology than CLIB. It contains approximately 100 concepts in total, whereas CLIB contains more than 1000, 147 of which are action classes. In DOLCE, events are called *occurrents*. Entity-event relations are denoted by the expression *participation*. DOLCE distinguishes between *temporary* and *constant* participation (and other types of participation as well), distinctions that are not present in CLIB. Similarly to CLIB, DOLCE was used in domain-specific applications. Borgo and Leitão, for instance, used DOLCE to model a manufacturing domain [5].

Other commonly used upper ontologies are: Basic Formal Ontology (BFO) [17, 29] containing 36 classes in total; General Formal Ontology (GFO) [18] containing 79 classes; or Suggested Upper Merged Ontology (SUMO) [23] containing 20,000 terms. As far as we know, there is no published research on guidelines for encoding knowledge described by natural language sentences, for any of these ontologies. However, we

believe that the method we describe in this paper is general enough to be applicable to these upper ontologies as well.

There are several specialized biological or biomedical ontologies currently in use. They generally tend to have a large number of concepts. Systems Biology Ontology (SBO) [14] is an ontology dedicated to a specific branch of biology. It incorporates the concept of *interaction*, which roughly corresponds to events in CLIB. The Gene Ontology (GO) [13] is designed to facilitate the description of gene products. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [31] is a much larger biomedical ontology, containing over 400,000 concepts. It is currently in use in different countries. There has been substantial research in revising and auditing this large ontology [25, 32, 28]. In contrast with the issues we discussed in relation to CLIB, the problems identified by this body of work concerned the *taxonomy* of SNOMED-CT. Some similarities with our approach are present however, such as a close collaboration between knowledge engineers and domain experts, and a need to address the mismatch between a common sense meaning of words and their usage in the ontology.

A different type of research with converging goals to ours is Proposition Bank (PropBank) [24] — "a corpus of text annotated with information about basic semantic propositions." The goal of PropBank is to define a methodology for mapping nouns in a sentence into *arguments* of the verb in that sentence. PropBank arguments correspond loosely to relations of CLIB, but a PropBank argument may reflect the meaning of one or more CLIB relations (e.g., Arg0 denotes both agents and experiencers). As a result, the task we address is much more difficult than the one of PropBank.

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 instance, the arguments specified for the verb *to move* are: (a) Arg0: mover (b) Arg1: moved (c) Arg2: destination. If the same noun (entity) plays more than one role in a sentence, only the argument with the highest rank is assigned. This solution could be used in our application as well, in order to prevent awkward translations into natural language when the same entity appears several times in a sentence.

A second resource used by annotators is a detailed set of guidelines provided by [4] for the mapping of nouns into arguments, with specific instructions for sentences with different syntactic structures (e.g., declarative sentences, questions, etc.). Our work also focuses on developing guidelines for a consistent assignment of entities to participant relations of events, but we operate at a higher level of abstraction. We do not look at sentences expressed in natural language directly; rather we assume that sentences are transformed into Universal Truths first.

7 Summary and Conclusions

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 some key aspects of creating a knowledge base from a biology textbook. Even though we used specific examples from our project, there are three broad lessons that are of interest to other projects using both manual and automated techniques for knowledge acquisition. These lessons are: (1) reformulating the

sentences so that their abstract structure is closer to the logical form to be acquired (2) use of a linguistically motivated upper ontology (3) use of a combination of syntactic and semantic guidelines to specify how ontological distinctions are expressed in language. We further hope that the three lessons at a general level, and the specifics of the guidelines that we presented, 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.

8 Acknowledgment

This work has been funded by Vulcan Inc. and SRI International.

Relation	Semantic Definition	Syntactic Definition	Biology Examples
agent	The entity that initiates, performs, or causes an event.	• the grammatical subject of a sentence in active voice • preposition: *by* (sentence in passive voice) (Assume that biological entities like protein, bacteria, etc., can be agents too.)	*A virus* **enters** a cell. A cell **is penetrated** *by a virus*.
object	The entity that is acted upon by an event; the main passive participant in the event.	• the grammatical object of a sentence in active voice • 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.	• the grammatical object of verbs like *to use*, *to consume*, etc. • preceded by: *using*	The Calvin cycle **uses** *the ATP and NADPH* to produce sugar. *Water* **is converted** to hydrogen. Chemicals **are transported**, *using energy*.
result	The entity that comes into existence as a result of an event.	• the grammatical object of verbs like *to produce*, *to create*, etc. • 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 transfered** *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*.
beneficiary	The entity that benefits from an event.	• preposition: *for*	
experiencer	The entity that experiences an event.	For a sentence containing a verb describing an emotional or psychological action: • the sentence subject (sentence in active voice) • preposition: *by* (sentence in passive voice)	*Plants* **sense** gravity and the direction of light. Gravity and the direction of light **are sensed** *by plants*.
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*.
away-from	The place away from which an event transpires, but not necessarily where the event starts.	• preposition: *away from*	The plasma membrane **pulls** *away from the wall*.
toward	The place toward which an event transpires, but not necessarily where the event ends.	• preposition: *toward*	Daughter chromosomes **move** *toward opposite ends of the cell*.
path	The place (or other entity) along or through which an entity moves.	• preposition: *across*, *along*, *through*	A protein **moves** into a cell *through a pore*.
site	The specific place of some effect of an event, as opposed to the locale of the event itself.	• preposition: *at*	The protein **binds** *at the groove* with the target molecules of bacterial walls.

Table 4: Summary of guidelines for mapping entities into slots.

Bibliography

[1] Eva Banik, Eric Kow, and Vinay K. Chaudhri. User-controlled, robust natural language generation from an evolving knowledge base. In *ENLG 2013 : 14th European Workshop on Natural Language Generation*, 2013.

[2] K. Barker, T. Copeck, S. Delisle, and S. Szpakowicz. Systematic construction of a versatile case system. *Journal of Natural Language Engineering*, 3(4):279–315, 1997.

[3] K. Barker, B. Porter, and P. Clark. A library of generic concepts for composing knowledge bases. In *First International Conference on Knowledge Capture*, 2001.

[4] Claire Bonial, Jena Hwang, Julia Bonn, Kathryn Conger, Olga Babko-Malaya, and Martha Palmer. English PropBank annotation guidelines, 2012.

[5] Stefano Borgo and Paulo Leitão. The role of foundational ontologies in manufacturing domain applications. *LNCS*, 2004.

[6] Stefano Borgo and Claudio Masolo. Foundational choices in DOLCE. In Steffen Staab and Ruder Studer, editors, *Handbook on Ontologies*. Springer, second edition, 2009.

[7] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *AAAI*, 2010.

[8] Vinay K Chaudhri, Britte Cheng, Adam Overholtzer, Jeremy Roschelle, Aaron Spaulding, Peter Clark, Mark Greaves, and Dave Gunning. Inquire Biology: A textbook that answers questions. *AI Magazine*, 34(3), September 2013.

[9] Vinay K. Chaudhri, Nikhil Dinesh, and Craig Heller. Conceptual models of structure and function. Technical report, SRI International, 2013.

[10] Vinay K. Chaudhri, Stijn Heymans, Michael Wessel, and Son Cao Tran. Query answering in object oriented knowledge bases in logic programming. In *Workshop on ASP and Other Computing Paradigms*, 2013.

[11] Vinay K. Chaudhri, Michael A. Wessel, and Stijn Heymans. KB_Bio_101: A challenge for OWL Reasoners. In *The OWL Reasoner Evaluation Workshop*, 2013.

[12] P. Clark, J. Thompson, K. Barker, B. Porter, V. Chaudhri, A. Rodriguez, J. Thomere, S. Mishra, Y. Gil, P. Hayes, and T. Reichherzer. Knowledge entry as the graphical assembly of components. In *First International Conference on Knowledge Capture*, 2001.

[13] The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet*, 25:25–29, 2000.

[14] M. Courtot, N. Juty, C. Knpfer, D. Waltemath, A. Zhukova, A. Drger, M. Dumontier, A. Finney, M. Golebiewski, J. Hastings, S. Hoops, S. Keating, D.B. Kell, S. Kerrien, J. Lawson, A. Lister, J. Lu, R. Machne, P. Mendes, M. Pocock, N. Rodriguez, A. Villeger, D.J. Wilkinson, S. Wimalaratne, C. Laibe, M. Hucka, and N. Le Novre. Model storage, exchange and integration. *Molecular Systems Biology*, 7(543), 2011.

[15] Nikhil Dinesh, Aravind K. Joshi, Insup Lee, and Oleg Sokolsky. Logic-based regulatory conformance checking. In *Monterey Workshop*, pages 147–160, 2007.
[16] Aldo Gangemi, Nicola Guarino, Claudio Masolo, and Alessandro Oltramari. Sweetening Wordnet with DOLCE. *AI Magazine*, 24(3):13–24, 2003.
[17] P. Grenon, B. Smith, and L. Goldberg. Applying BFO in the biomedical domain. *Health Technology and Informatics*, 102:20–38, 2004.
[18] Heinrich Herre, Barbara Heller, Patryk Burek, Robert Hoehndorf, Frank Loebe, and Hannes Michalek. General formal ontology (GFO): A foundational ontology integrating objects and processes. `http://www.onto-med.de/ontologies/gfo/`, 2013.
[19] Graeme Hirst. Ontology and the lexicon. In *Handbook on ontologies*, pages 269–292. Springer, 2009.
[20] S. M. Lloyd, editor. *Roget's Thesaurus*. Longman, 1982.
[21] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics, 2003.
[22] George A. Miller and Christiane Fellbaum. Wordnet then and now. *Language Resources and Evaluation*, 41(2):209–214, 2007.
[23] Ian Niles and Adam Pease. Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems - Volume 2001*, FOIS '01, pages 2–9, New York, NY, USA, 2001. ACM.
[24] Martha Palmer, Dan Gildea, and Paul Kingsbury. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics Journal*, 31(1), 2005.
[25] Alan Rector, Luigi Iannone, and Robert Stevens. Quality assurance of the content of a large DL-based terminology using mixed lexical and semantic criteria: experience with SNOMED CT. In *Proceedings of the sixth international conference on Knowledge capture*, K-CAP '11, pages 57–64, New York, NY, USA, 2011. ACM.
[26] Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, and Robert B. Jackson. *Campbell biology*. Benjamin Cummings imprint of Pearson, Boston, 2011.
[27] Brian C. Smith. *Reflection and Semantics in a Procedural Language*. PhD thesis, Massachusetts Institute of Technology, 1982.
[28] Kent A. Spackman and Guillermo Reynoso. Examining SNOMED from the perspective of formal ontological principles: Some preliminary analysis and observations. In *KR-MED*, pages 72–80, 2004.
[29] A. D. Spear. Ontology for the twenty first century: An introduction with recommendations. `http://www.ifomis.org/bfo/documents/manual.pdf`, 2006.
[30] D. Summers, editor. *Longman Dictionary of Contemporary English*. Longman, 1987.
[31] Amy Y Wang, Jeremiah H Sable, and Kent A Spackman. The SNOMED clinical terms development process: refinement and analysis of content. *Proc AMIA Symp*, pages 845–9, 2002.
[32] Yue Wang, Michael Halper, Hua Min, Yehoshua Perl, Yan Chen, Kent A. Spackman, All Communications To, and Yehoshua Perl. Structural methodologies for auditing SNOMED.