Using Scripts to help in Biomedical Text Interpretation

Working Note 30

Peter Clark (peter.e.clark@boeing.com), Boeing Research and Technology, 2009

Introduction

This short note speculates on the use of world knowledge to help interpret a short paragraph of biomedical text about transcytosis (transport across a cell). The ultimate goal is to create a simple representation of the transcytosis process from the text (either automatically or semi-automatically). The challenges are formidable, as the process involves several steps that are often expressed opaquely (or even omitted) in the text, and the whole process is described over several non-contiguous sentences.

The approach illustrated here is to use a more general model of transportation, essentially a script (Schank and Abelson, 1977), to guide, constrain, and elaborate interpretation of the text. This is not a new idea -- there are numerous examples in the literature of using scripts for interpretation dating back to the '70s, e.g., FRUMP (DeJong, 1979), SAM (Cullingford, 1977), with more recent work in the Message Understanding Competitions (e.g., Riloff, 1991), by Wagner et al. (2009), and by several other authors (described in Ram and Moorman, 1991). Rather, this note explores applying those ideas in the specific context of a biomedical text.

Hunter et al. have previously used sentence-level semantic patterns for biomedical information extraction (Hunter et al, 2008). The semantic patterns include semantic constraints on arguments, blocking some nonsensical interpretations. For example, a pattern might be

> *transported-entity* "translocation" ("from" *det*? *transport-origin)*?
> ("to" *det*? *tranport-destination*)?

where *transported-entity, transport-origin*, and *transport-desination* are constrained to be, respectively, of type Protein or Molecular-Complex, Cellular-Component, and Cellular-Component (Hunter et al., 2008). Semantic patterns and their constraints serve to guide interpretation at the sentence level. However, although these patterns may extract elements of a transportation process, there is at present no mechanism for assembling an overall representation of the whole process automatically or semi-automatically. This note explores that goal using scripts. One might view this as a generalization of semantic patterns, in that a script can be viewed as a paragraph-sized semantic pattern.

The Scenario

The particular scenario we look at here is a short paragraph from (Alberts et al, 2007) describing transportation of a protein from the plasma membrane on one side of the cell to the other:

> In contrast to clathrin-coated and COPI- or COPII-coated vesicles, caveolae are thought to invaginate and collect cargo proteins by virtue of the lipid composition of the calveolar membrane, rather than by the assembly of a cytosolic protein coat. Caveolins may stabilize these raft domains, into

> which certain plasma membrane proteins partition. Caveolae pinch off from the plasma membrane using dynamin, and they deliver their contents either to an endosomelike compartment (called a caveosome) or to the plasma membrane on the opposite side of a polarized cell (in a process called transcytosis, which we discuss later).

Or, restating (and also slightly elaborating) this in non-biomedical terminology (to the best of my ability): The paragraph describes how proteins are transported across the cell. First the protein is collected in little pits ("caves", or caveolae) in the plasma membrane, then the pits (caveolae) close and bud off (a bit like a bubble), then the bubbles (now called vesicles) move across the cell, then they finally fuse with the plasma membrane on the other side of the cell and open up again to release their contents. In reality, the bubbles (vesicles) do not traverse the whole cell, rather they only travel to intermediate way-stations (endosomes, including caveosomes). At endosomes they fuse and release their contents into the endosome, then later new "bubbles" (vesicles) form from and leave the endosome to carry the cargo further across the cell.

Because this is a transportation process, there are numerous important constraints that hold, based on general properties of transportation processes. For example:

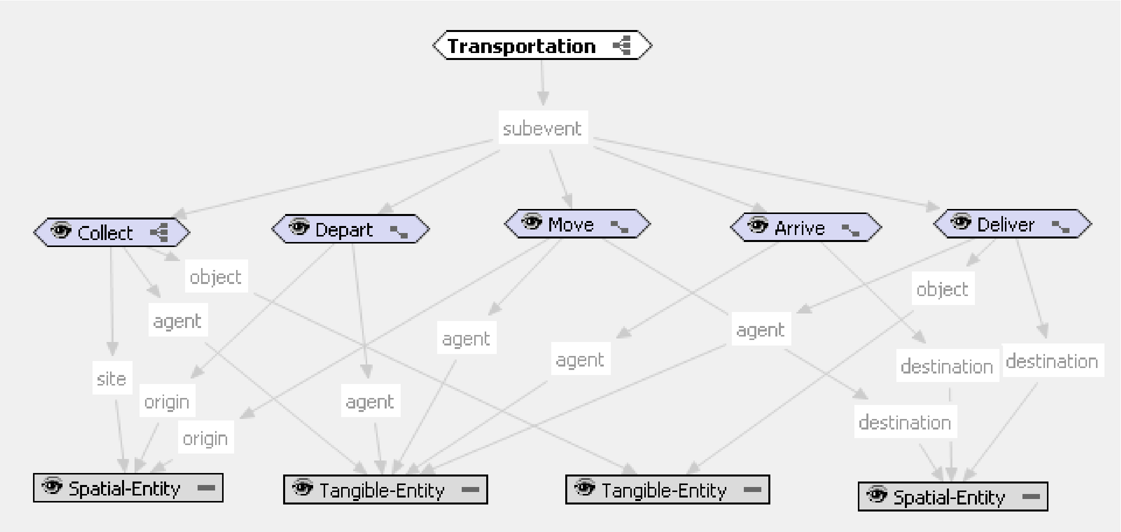
- the cargo collected = the cargo moved = the cargo released
- the thing moving the cargo = the thing releasing the cargo
- the destination of the movement = the place where the cargo is later released

Knowing these constraints is essential to understanding the original text, because the original text is often is vague, or uses alternative words, or does not mention some of these details. The constraints can thus serve to fill in the gaps, identify coreference, and filter out irrelevant text. For example, the original text refers to the cargo as both "cargo proteins" and "contents"; knowledge from the script suggests these are alternative descriptions of the same object, and thus helps to align the different descriptions. Similarly, the original text does not mention the movement across the cell itself - instead it describes just the collection and delivery of the cargo, and the movement is left "obvious" to the reader. Again, the script can help fill in this important, implicit step, not only hypothesizing the movement event but also the source, destination, and object that was moved.

A simple representation (script) of transportation might look, with some context-specific biomedical synonyms in parentheses, as follows:

> X **transports** ("transcytosis") Y from A to B:
>
> > X **collects** Y at A | Y **enters** X at A
> > X (and thus Y) **departs** ("pinches off", "buds") from A
> > X (and thus Y) **moves** ("traverses", "transports") from A to B
> > X (and thus Y) **arrives** at B
> > X **delivers** ("deposits") Y to B | Y **exits** X at B

where X is a vehicle ("vesicle", "container"), Y is a cargo ("contents"), and A and B are locations. This general model of transportation can be sketched graphically as shown on the next page.



How do the elements of this script manifest themselves in the text? Capitalizing script elements in the paragraph (as an automatic extraction system might feasibly do), we see:

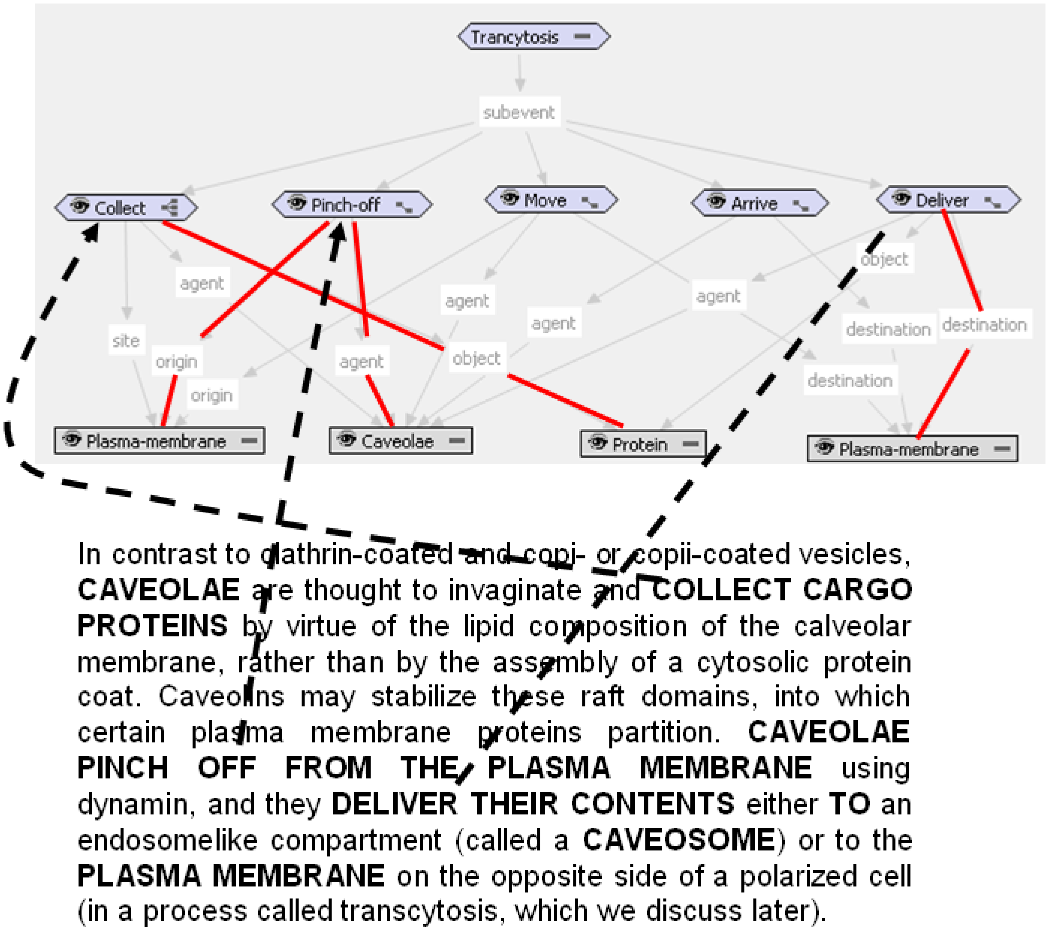
> In contrast to clathrin-coated and copi- or copii-coated vesicles, **CAVEOLAE** are thought to invaginate and **COLLECT CARGO PROTEINS** by virtue of the lipid composition of the calveolar membrane, rather than by the assembly of a cytosolic protein coat. Caveolins may stabilize these raft domains, into which certain plasma membrane proteins partition. **CAVEOLAE PINCH OFF FROM THE PLASMA MEMBRANE** using dynamin, and they **DELIVER THEIR CONTENTS** either **TO** an endosomelike compartment (called a **CAVEOSOME**) or to the **PLASMA MEMBRANE** on the opposite side of a polarized cell (in a process called transcytosis, which we discuss later).

The relevant, sentence-based facts that can by extracted from the text (best case) are as follows:

- Caveolae collect cargo proteins.
- Caveolae pinch off from the plasma membrane.
- Caveolae deliver to the caveosome/plasma membrane.

This might be the typical output of a sentence-based extraction system. However, a script-based extraction system could feasibly produce considerably more, where the text is used to instantiate the script, resulting in the script containing the knowledge that:

- Caveolae collect cargo proteins at the plasma membrane.
- Caveolae + cargo proteins pinch off from the plasma membrane.
- Caveolae + cargo proteins move from the plasma membrane to the caveosome/plasma membrane.
- Caveolae + cargo proteins arrive at the caveosome/plasma membrane.
- Caveolae deliver the cargo proteins to the caveosome/plasma membrane.



In contrast to clathrin-coated and copi- or copii-coated vesicles, **CAVEOLAE** are thought to invaginate and **COLLECT CARGO PROTEINS** by virtue of the lipid composition of the calveolar membrane, rather than by the assembly of a cytosolic protein coat. Caveolins may stabilize these raft domains, into which certain plasma membrane proteins partition. **CAVEOLAE PINCH OFF FROM THE PLASMA MEMBRANE** using dynamin, and they **DELIVER THEIR CONTENTS** either **TO** an endosomelike compartment (called a **CAVEOSOME**) or to the **PLASMA MEMBRANE** on the opposite side of a polarized cell (in a process called transcytosis, which we discuss later).

Figure 1: Only a small portion of the full script (highlighted in the graph) is stated explicitly in the text. The script can then help to fill in the unstated information, from the general knowledge of transportation events that it embodies.

The most interesting thing to note is the large amount of information in the text that is unstated (implicit), but which the script can supply, as illustrated in the sentence-based vs. script-based extractions shown on the previous page and pictorially in Figure 1. For example, "caveolae collect cargo proteins" does not specify the location of that collection, but we can infer this (using the script) because the text later says "caveolae pinch off (depart) from the plasma membrane", and the script states that the point of departure = the point of collection (= A). Similarly the text does not state what is delivered ("deliver their contents"), but again the script tells us the thing delivered = the thing collected = the cargo proteins (= Y). The movement itself across the cell is completely unmentioned, although the script tells us this must have occurred.

In addition, the script provides important information about *what* to look for in the text. The sentence-based extractions shown earlier are essentially disconnected fragments -

they may miss out important facts, and they may include irrelevant facts, and they are unordered. In contrast, the script provides strong expectations: Once the text has been identified as describing transportation, the script expects to see mention of collect, depart, move, arrive, and deliver events, and in that order. This can dramatically constrain the search and help bias interpretation of ambiguous and/or unrecognized terminology. There is an important symbiotic relationship going on: The script tells the extraction system what to look for, while the extraction system instantiates the script and hence refines further expectations of what to look for.

Another point of interest and complication is a certain fluidity in terminology in the original text. The sample text talks about the caveolae pinching off and later delivering their contents. However, strictly speaking, once they have pinched off they become "vesicles" ("caveolae" only refers to the original pits doing the collection), and other texts will (more correctly) use the word "vesicle" rather than "caveolae", e.g., "transport vesicles rapidly bud and deliver large numbers of glucose transporters". Again, the script can help here: although the terminology changes, the script states that the collecting and transporting object must be the same (i.e., the collector = the transporter), as a general property of transportation. As a result, the fact that these two objects are the “same” can be inferred, even though the name changes during the process.

Architecture

Using scripts to help guide interpretation is an extension of, rather than alternative to, sentence-based extraction. In particular, looking for larger script-like structures still requires sentence-based extraction of scriptal components. One could envisage a system which took as input possible script fragments, extracted with OpenDMAP-style techniques (Hunter at al., 2008), and used them to assemble a script. Ultimately one would also like feedback the other way also, where a partially constructed script can send additional guidance to the sentence-based extractors, e.g., additional synonyms to look for in the text.

Knowledge Representation

A final point of interest is that the script provides a "compositional representation" of concepts, i.e., the basic entities involved in the process are represented as combinations of more primitive entities. For example, the instantiated script includes a representation of collecting and pinching off as structures which look:

collect::
agent: **caveoloe**
object: **protein**
location: **plasma membrane**

and

pinch off::
agent: **caveolae**
source: **plasma membrane**

While one might automatically reify these, for convenience, as long-named concepts like

"collection of cargo protein by caveolae from the plasma membrane"
and "caveoloae pinching off from the plasma membrane"

the structures provide a way of representing the *meaning* of those names. A knowledge representation capable of manipulating such structures removes the need to create such long names in the first place (although sometimes it is convenient to do so), and thus can avoid the otherwise endless enumeration that might ensue. In addition, it can allow users to specify constraints on what are are valid compositions, both at the atomic level (e.g., what are valid fillers of slots) and the structure level (e.g., in the script shown, that the thing collected = the thing transported = the thing delivered), similar to the notion of "sanctioning" in GALEN (Rector et al, 1994).

Use of Additional Knowledge

In this note, we have only explored use of one piece of domain-general knowledge (about transportation) to guide interpretation. There is clearly substantially more domain-general and biology-specific knowledge that could similarly be employed to guide interpretation.

Summary and Discussion

We have sketched out how a larger knowledge structure, here a script, can serve to guide and elaborate interpretation of text. In addition to providing additional constraints on the interpretation, the script can also serve to fill in gaps and suggest coreferences that were otherwise omitted or would be hard to infer in the original text.

The use of scripts is not a new idea. Rather, this note explores applying old script ideas dating back to the '70s to the biomedical arena. In addition, it is now well known that the use of scripts is not a panacea: scripts have not "solved" information extraction, and there are several well-known problems that would need to be addressed in any larger-scale exploration of this idea, e.g., see (Clark, 2008), including:

- There is rarely a single script for a process. Rather, a process typically involves composition of several scripts (e.g., one of several scripts may be used to perform a step in some other script).
- Only a small fraction of a script is explicitly stated in text, i.e., the script's expression in language is sparse. This was illustrated in the example earlier, and that level of sparcity is typical of language in general. The sparcity means that it is critical to extract the linguistic information that *is* explicitly stated, and any failure of the extraction can leave critical gaps in the final representation.
- Language itself is highly complex, meaning the basic process of linguistic analysis, information extraction, and alignment of information with knowledge remain challenging.

Despite these problems, there has been dramatic progress in language processing, information extraction, and language interpretation technologies since the 1970's. There is a good case to make that these advances will make these earlier problems more surmountable, potentially dramatically improving knowledge-guided information extraction from text.

References

Alberts, B., Wilson, J., Johnson, A., Lewis, J. Raff, M., Roberts, M. “Molecular Biology of the Cell”, 2007.

Clark, P. "Do Scripts solve NLP?", Working Note 28, 2008. http://www.cs.utexas.edu/users/pclark/working_notes/

Cullingford, R., "Controlling Inference in Story Understanding", IJCAI'77.

DeJong, G. "Prediction and Substantiation: A New Approach to Natural Language Processing" Cognitive Science 3(for any task) July 1979 , pp251-271.

Hunter, L., Lu, Z., Firby, J., Baumgartner, W., Johnson, H., Ogren, P., Cohen, K. "OpenDMAP: An open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression". In BMC Bioinformatics, 2008, 9:78.

Ram, A., Moorman, K. “Understanding Language Understanding”, MA:MIT, 1991.

Rector, A., Gangemi, A., Galeazzi, E., Glowinski, A., Rossi-Mori, A., “The GALEN CORE Model Schemata for Anatomy”, in Proc Medical Informatics Europe (MIE’94), pp186-189, 1994.

Riloff, E. "Information Extraction as a Stepping Stone toward Story Understanding", in Understanding Language Understanding, Ed. A. Ram, K. Morroman. 1991.

Schank, R., Abelson, R., "Scripts, Plans, Goals, and Understanding". NJ:Erlbaum, 1977.

Wagner, E., Liu, J., Birnbaum, L., Forbus, K. “Rich Interfaces for Reading News on the Web”, in IUI’2009.