Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Understanding Descriptions of Visual Scenes Using Graph Grammars

Daniel Bauer
Columbia University
1214 Amsterdam Avenue, 450 Computer Science Building
New York, NY 10025
bauer@cs.columbia.edu

Teaching computers to understand the meaning of natural language text has long been an important goal for Artificial Intelligence. The focus of my work is on the interpretation of descriptions of visual scenes such as *'A man is sitting on a chair and using the computer'*. One application of this research is the automatic generation of 3D scenes (Coyne and Sproat 2001), such as the one in Figure 1 c). Text-to-scene generation systems provide a way for non-artists to create graphical content and have wide-ranging applications in communication, entertainment, and education.

The formal meaning representations in today's natural language processing systems are usually limited to basic predicate-argument structure and coarse word sense. Such representations do not support inference and are not sufficiently detailed to visualize a scene.

In my thesis I am developing techniques for semantic parsing into a new type of meaning representation encoded as directed graphs. Graphs conveniently capture coreference and the hierarchical nature of meaning. My meaning representations contain two or more levels of granularity. The graph directly derived from the input text (the *high-level* representation) describes functional aspects of the scene (*who does what to whom*, Figure 1 a). It can be rewritten into a *low-level* graph that contains concepts and relations that are more basic (Figure 1 b) and eventually into conceptual primitives. For visual scenes, these low-level graphs express the basic spatial relations between objects.

This meaning representation scheme is based on two powerful ideas in natural language understanding: decomposing word meaning into conceptual primitives to support inference (Schank 1972) and describing word meaning not as isolated fragments but as part of a larger conceptual frame. In particular I build on the frame semantic theory by (Fillmore 1982) and its implementation in the FrameNet lexical resource (Fillmore, Johnson, and Petruck 2003). Fillmore's frame semantics focuses on valence patterns of a lexical item as the link between syntactic realization and elements of the conceptual frame surrounding it. There are a number of systems using FrameNet as training data to automatically annotate frame semantic structure on text (e.g. Das et al. 2010). FrameNet representations, however, are shallow: frames do not contain any internal structure and frame elements are associated with phrases, not with semantic objects.

In (Coyne, Bauer, and Rambow 2011) we extended frame semantics with conceptual decomposition to represent visual scenes. We also constructed VigNet, a knowledge base and lexical resource that builds on FrameNet and provides decomposition of frames into primitives based on their frame elements. I propose to encode these representations as graphs. Every node is a frame instance. Outgoing edges describe the internal structure of a frame, including its frame elements (see Figure 1). These graphs can also be interpreted as typed attribute value matrices.

To convert input text into a high-level meaning representation and to rewrite this graph into the low-level representation I propose to use Hyperedge Replacement Grammars (HRG, Drewes, Habel, and Kreowski 1997). These grammars generate languages of graphs in much the same way that context free string grammars generate string languages. To derive a graph HRGs repeatedly substitute nonterminal (hyper)-edges in a graph with larger graph fragments. In my encoding, the graph fragments correspond to the internal structure of a frame.

In previous work (Jones et al. 2012) we extended HRG to a synchronous grammar formalism (SHRG) that pairs each HRG rule with a rule in a string grammar. To convert a sentence into possible high-level graphs we first parse the input sentence with the string side of the grammar and then use the derivation forest to re-assemble a set of graphs. We use SHRG for semantic parsing and generation in a semantics-based machine translation system. The shallow graph-based meaning representation employed there differs from the one proposed in my thesis. Most significantly my graphs can be rewritten into a low-level representation as outlined above.

At this time I have only experimented with toy grammars that can parse a limited number of scene descriptions. I am working on constructing grammars automatically from VigNet.

I will also attempt to induce HRGs from scene descriptions paired with their meaning representation. Previous research exists on learning semantic parsers from data (e.g. Wong and Mooney 2006), but little work has been done for graph-based meaning representations. In (Jones et al. 2012) we developed methods for the automatic acquisition of SHRGs. For the more difficult problem of acquiring conversion rules from high-level into low-level scene represen-

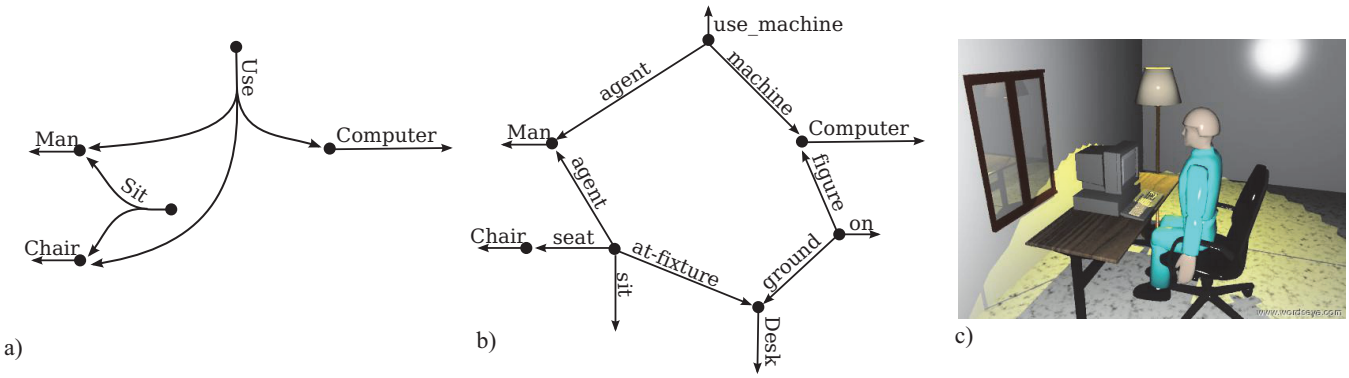

Figure 1: High-level (a) and low-level (b) semantic representation and a corresponding 3D-scene (c).

tations I am planning to use clustering algorithms.

Such learning methods need a suitable data set. I am developing an annotation tool to create semantic annotations of scene descriptions by hand. Another source of data is the WordsEye system (Coyne and Sproat 2001), which does not support high-level language input. Using the system, we have collected a corpus of 13,000 low-level textual descriptions with rule-generated scene representations. I am collecting additional data from a more restricted domain, such as domestic scenes. In a first experiment I will attempt to induce grammars for semantic parsing into low-level graphs that mimic WordsEye's rules. I will also employ crowdsourcing to collect high-level text describing the existing scenes.

Not all graphs generated by a grammar are valid meaning representations. A decomposition into a low-level graph may only be permitted if the objects involved are of a certain type (a scene for *'use computer'* looks different from a scene for *'use cell phone'*). Furthermore the decomposition of two different predicates may be mutually exclusive because of spatial impossibilities or conflicts on a shared resource. Such inconsistencies require global reasoning to resolve. In addition some representations may be preferred over others.

My thesis will explore different approaches to enforcing soft and hard constraints. One possibility is to convert the set of possible graphs into a constraint satisfaction problem and to use an off-the-shelf constraint solver. I am also exploring Lagrangian relaxation methods (Martins et al. 2011) to jointly unpack the derivation forest while enforcing constraints and maximizing the score for the resulting meaning representation.

To evaluate my efforts I propose to use graph similarity measures on my existing and new data sets. Finally I will generate actual 3D-scenes from low-level representations using the graphical processing engine of WordsEye. for a human-based end-to-end evaluation.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No.IIS-0904361.

F 2013	Annotation / data collection. Parsing with VigNet.
S 2014	Automatic grammar acquisition.
F 2014	Constraint satisfaction.
S 2015	Final experiments. Write-up.
F 2015	Thesis Defense.

Table 1: Timeline for work on thesis components.

References

Coyne, B., and Sproat, R. 2001. Wordseye: an automatic text-to-scene conversion system. In *Proceedings of SIGGRAPH*.

Coyne, B.; Bauer, D.; and Rambow, O. 2011. Vignet: Grounding language in graphics using frame semantics. In *ACL Workshop on Relational Semantics (RELMS)*.

Das, D.; Schneider, N.; Chen, D.; and Smith, N. 2010. Probabilistic frame-semantic parsing. In *Proceedings of NAACL*.

Drewes, F.; Habel, A.; and Kreowski, H. 1997. Hyperedge replacement graph grammars. In Rozenberg, G., ed., *Handbook of Graph Grammars and Computing by Graph Transformation*. World Scientific. 95–162.

Fillmore, C.; Johnson, C.; and Petruck, M. 2003. Background to FrameNet. *International Journal of Lexicography* 16(3):235–250.

Fillmore, C. 1982. Frame semantics. In of Korea, L. S., ed., *Linguistics in the Morning Calm*. Seoul: Hanshin. 111–137.

Jones, B.; Andreas*, J.; Bauer*, D.; Hermann*, K. M.; and Knight, K. 2012. Semantics-based machine translation with hyperedge replacement grammars. In *Proceedings of COLING*. *first authorship shared.

Martins, A. F. T.; Smith, N. A.; Aguiar, P. M. Q.; and Figueiredo, M. A. T. 2011. Dual decomposition with many overlapping components. In *Proceedings of EMNLP*.

Schank, R. 1972. Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology* 10(3):552–631.

Wong, Y. W., and Mooney, R. J. 2006. Learning for semantic parsing with statistical machine translation. In *Proceedings of NAACL*.