Building Multimodal Simulations for Natural Language

ACL 2017 Tutorial

James Pustejovsky **Nikhil Krishnaswamy**

1 Tutorial Description

In this tutorial, we introduce a computational framework and modeling language (VoxML) for composing multimodal simulations of natural language expressions within a 3D simulation environment (VoxSim). We demonstrate how to construct *voxemes*, which are visual object representations of linguistic entities. We also show how to compose events and actions over these objects, within a restricted domain of dynamics. This gives us the building blocks to simulate narratives of multiple events or participate in a multimodal dialogue with synthetic agents in the simulation environment. To our knowledge, this is the first time such material has been presented as a tutorial within the CL community.

This will be of relevance to students and researchers interested in modeling actionable language, natural language communication with agents and robots, spatial and temporal constraint solving through language, referring expression generation, embodied cognition, as well as minimal model creation.

Multimodal simulation of language, particularly motion expressions, brings together a number of existing lines of research from the computational linguistic, semantics, robotics, and formal logic communities, including action and event representation (Di Eugenio, 1991), modeling gestural correlates to NL expressions (Kipp et al., 2007; Neff et al., 2008), and action event modeling (Kipper and Palmer, 2000; Yang et al., 2015). We combine an approach to event modeling with a scene generation approach akin to those found in work by (Coyne and Sproat, 2001; Siskind, 2011; Chang et al., 2015). Mapping natural language expressions through a formal model and a dynamic logic interpretation into a visualization of the event described provides an environment for grounding concepts and referring expressions that is interpretable by both a computer and a human user. This opens a variety of avenues for humans to communicate with computerized agents and robots, as in (Matuszek et al., 2013; Lauria et al., 2001), (Forbes et al., 2015), and (Deits et al., 2013; Walter et al., 2013; Tellex et al., 2014). Simulation and automatic visualization of events from natural language descriptions and supplementary modalities, such as gestures, allows humans to use their native capabilities as linguistic and visual interpreters to collaborate on tasks with an artificial agent or to put semantic intuitions to the test in an environment where user and agent share a common context.

In previous work (Pustejovsky and Krishnaswamy, 2014; Pustejovsky, 2013a), we introduced a method for modeling natural language expressions within a 3D simulation environment built on top of the game development platform Unity (Goldstone, 2009). The goal of that work was to evaluate, through explicit visualizations of linguistic input, the semantic presuppositions inherent in the different lexical choices of an utterance. This work led to two additional lines of research: an explicit encoding for how an object is itself situated relative to its environment; and an operational characterization of how an object changes its location or how an agent acts on an object over time, e.g., its affordance structure. The former has developed into a semantic notion of situational context, called a *habitat* (Pustejovsky, 2013a; McDonald and Pustejovsky, 2014), while the latter is addressed by dynamic interpretations of event structure (Pustejovsky and Moszkowicz, 2011; Pustejovsky and Krishnaswamy, 2016b; Pustejovsky, 2013b).

The requirements on building a visual simulation from language include several components. We require a rich type system for lexical items and their composition, as well as a language for modeling the dynamics of events, based on Generative Lexicon (GL). Further, a minimal embedding space (MES) for the simulation must be determined. This is the 3D region within which the state is configured or the event unfolds. Object-based attributes for participants in a situation or event also need to be specified;

e.g., orientation, relative size, default position or pose, etc. The simulation establishes an epistemic condition on the object and event rendering, imposing an implicit point of view (POV). Finally, there must be some sort of agent-dependent embodiment; this determines the relative scaling of an agent and its event participants and their surroundings, as it engages in the environment.

In order to construct a robust simulation from linguistic input, an event and its participants must be embedded within an appropriate minimal embedding space. This must sufficiently enclose the event localization, while optionally including space enough for a frame of reference for the event (the viewer's perspective).

We first describe the formal multimodal foundations for the modeling language, VoxML, which creates a *minimal simulation* from the linguistic input interpreted by the multimodal language, DITL. We then describe VoxSim, the compositional modeling and simulation environment, which maps the minimal VoxML model of the linguistic utterance to a simulation in Unity. This knowledge includes specification of object affordances, e.g., what actions are possible or enabled by use an object.

VoxML (Pustejovsky and Krishnaswamy, 2016b; Pustejovsky and Krishnaswamy, 2016a) encodes semantic knowledge of real-world objects represented as 3D models, and of events and attributes related to and enacted over these objects. VoxML goes beyond the limitations of existing 3D visual markup languages by allowing for the encoding of a broad range of semantic knowledge that can be exploited by a simulation platform such as VoxSim.

VoxSim (Krishnaswamy and Pustejovsky, 2016a; Krishnaswamy and Pustejovsky, 2016b) uses object and event semantic knowledge to generate animated scenes in real time without a complex animation interface. It uses the Unity game engine for graphics and I/O processing and takes as input a simple natural language utterance. The parsed utterance is semantically interpreted and transformed into a hybrid dynamic logic representation (DITL), and used to generate a minimal simulation of the event when composed with VoxML knowledge. 3D assets and VoxML-modeled nominal objects and events are created with other Unity-based tools, and VoxSim uses the entirety of the composed information to render a visualization of the described event.

The tutorial participants will learn how to build simulatable objects, compose dynamic event structures, and simulate the events running over the objects. The toolkit consists of object and program (event) composers and the runtime environment, which allows for the user to directly manipulate the objects, or interact with synthetic agents in VoxSim. As a result of this tutorial, the student will acquire the following skill set: take a novel object geometry from a library and model it in VoxML; apply existing library behaviors (actions or events) to the new VoxML object; model attributes of new objects as well as introduce novel attributes; model novel behaviors over objects.

The tutorial modules will be conducted within a build image of the software. Access to libraries will be provided by the instructors. No knowledge of 3D modeling or the Unity platform will be required.

2 Tutorial Outline

1. Introduction to Multimodal Semantics (1 hour)
   (a) Generative Lexicon Types and Habitat Theory:
   semantic encoding of how objects are situated and embodied.
   (b) Computational Theory of Affordances:
   extensions to Qualia Structure for encoding how an object can be manipulated and used.
   (c) Dynamic Logic for Event Structures:
   encoding how objects change over the course of an event.

2. Overview of VoxSim (.5 hours)
   (a) Module 1: Architecture and Program Flow:
   Natural language utterance is parsed (ClearNLP or other dependency parsers); output is interpreted in DITL (Dynamic Interval Temporal Logic); extracted entities are linked to objects in the simulator DB, extracted verbs are linked to programs; result is rendering in VoxSim.

(b) Module 2: Object Modeling:
the movement of objects independent of an agent

(c) Module 3: Action and Gesture Modeling:
motion of an agent independent of an object or caused movement, as well as creation of gestures

(d) Module 4: Event Modeling: integration of Object and Action Models:
Attaching an object to the rigged and animated skeleton of an agent as a surrogate for caused motion.

3. Creating Simulations: Modeling Novel Content (1.5 hours)

(a) Activity 1: Voxeme Modeling from 3D Geometry Library:
taking a 3D model of an object and augmenting it with semantic markup with VoxML, in order to provide semantic grounding in a model

(b) Activity 2: Behavior Attachment to a Voxeme (adding affordances):
executing afforded behaviors over newly created objects, in order to characterize the use, purpose, or manipulability of an object.

(c) Activity 3: Adding Discriminating Attributes to Voxemes:
Providing further object constraints in order to discriminate between entities in an environment.

(d) Activity 4: Creating Novel Behavior:
Composing existing actions into new behaviors (patterns of activities under constraints), and executing them over the objects from the Voxeme library.

3 Instructors

James Pustejovsky

Postal Address
Computer Science Department
Brandeis University
415 South St.
Waltham, MA 02453 USA
Phone: +1-617-301-2913
E-mail: jamesp@cs.brandeis.edu

Pustejovsky holds the TJX Feldberg Chair in Computer Science at Brandeis University, where he directs the Lab for Linguistics and Computation, and chairs both the Program in Linguistics and the Computational Linguistics Graduate Program. He conducts research in computational linguistics, lexical semantics, temporal and spatial reasoning, and language simulations. He has authored numerous books, including *Generative Lexicon*, MIT, 1995; *Semantics and the Lexicon*, Springer, 1993; *The Problem of Polysemy*, CUP, 1996 (with B. Boguraev); *The Language of Time*, OUP, 2005 (with I. Mani and R. Gaizauskas), *Interpreting Motion: Grounded Representations for Spatial Language*, OUP, 2012 (with I. Mani), and *Natural Language Annotation for Machine Learning*, O'Reilly, 2012 (with A. Stubbs); *Recent Advances in Generative Lexicon Theory*, Springer, 2013; *Handbook of Linguistic Annotation*, Springer, 2016, (edited with N. Ide); *A Guide to Generative Lexicon Theory*, OUP, 2017 (with E. Jezek).

Nikhil Krishnaswamy

Postal Address
Computer Science Department
Brandeis University
415 South St.
Waltham, MA 02453 USA
Phone : +1-614-592-4595

E-mail: nkrishna@brandeis.edu

Krishnaswamy is a Ph.D. candidate in Computer Science at Brandeis University, defending his thesis in Summer 2017. He has 8 years of experience in the software development, gaming and defense industries, building immersive training simulators for government and industry clients, and has a master's degree in Computational Linguistics. His research interests include the computational modeling of lexical semantics, spatial reasoning, human-computer communication, natural language understanding, and modal logic. Particular areas of expertise include procedural simulation generation and 3D visualization of motion events.

References

Angel Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D. Manning. 2015. Text to 3d scene generation with rich lexical grounding. *arXiv preprint arXiv:1505.06289*.

Bob Coyne and Richard Sproat. 2001. Wordseye: an automatic text-to-scene conversion system. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 487–496. ACM.

Robin Deits, Stefanie Tellex, Pratiksha Thaker, Dimitar Simeonov, Thomas Kollar, and Nicholas Roy. 2013. Clarifying commands with information-theoretic human-robot dialog. *Journal of Human-Robot Interaction*, 2(2):58–79.

Barbara Di Eugenio. 1991. Action representation for NL instructions. In *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, pages 333–334. Association for Computational Linguistics.

Maxwell Forbes, Rajesh PN Rao, Luke Zettlemoyer, and Maya Cakmak. 2015. Robot programming by demonstration with situated spatial language understanding. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2014–2020. IEEE.

Will Goldstone. 2009. *Unity Game Development Essentials*. Packt Publishing Ltd.

Michael Kipp, Michael Neff, Kerstin H Kipp, and Irene Albrecht. 2007. Towards natural gesture synthesis: Evaluating gesture units in a data-driven approach to gesture synthesis. In *International Workshop on Intelligent Virtual Agents*, pages 15–28. Springer.

Karin Kipper and Martha Palmer. 2000. Representation of actions as an interlingua. In *Proceedings of the 2000 NAACL-ANLP Workshop on Applied interlinguas: practical applications of interlingual approaches to NLP-Volume 2*, pages 12–17. Association for Computational Linguistics.

Nikhil Krishnaswamy and James Pustejovsky. 2016a. Multimodal semantic simulations of linguistically underspecified motion events. *Proceedings of Spatial Cognition*.

Nikhil Krishnaswamy and James Pustejovsky. 2016b. VoxSim: A visual platform for modeling motion language. *Proceedings of COLING System Demonstrations*.

Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, Johan Bos, and Ewan Klein. 2001. Personal robot training via natural-language instructions. *IEEE Intelligent systems*, 16(3):38–45.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2013. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer.

David McDonald and James Pustejovsky. 2014. On the representation of inferences and their lexicalization. In *Advances in Cognitive Systems*, volume 3.

Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. 2008. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics (TOG)*, 27(1):5.

James Pustejovsky and Nikhil Krishnaswamy. 2014. Generating simulations of motion events from verbal descriptions. *Lexical and Computational Semantics (* SEM 2014)*, page 99.

J. Pustejovsky and N. Krishnaswamy. 2016a. Constructing visualizing events: Simulating meaning in language. In *The 38th Cognitive Science Conference COGSCI 2016*, Philadelphia, PA.

James Pustejovsky and Nikhil Krishnaswamy. 2016b. VoxML: A visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, May. European Language Resources Association (ELRA).

James Pustejovsky and Jessica Moszkowicz. 2011. The qualitative spatial dynamics of motion. *The Journal of Spatial Cognition and Computation*.

James Pustejovsky. 2013a. Dynamic event structure and habitat theory. In *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, pages 1–10. ACL.

James Pustejovsky. 2013b. Where things happen: On the semantics of event localization. In *Proceedings of ISA-9: International Workshop on Semantic Annotation*.

Jeffrey Mark Siskind. 2011. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *arXiv preprint arXiv:1106.0256*.

Stefanie Tellex, Pratiksha Thaker, Joshua Joseph, and Nicholas Roy. 2014. Learning perceptually grounded word meanings from unaligned parallel data. *Machine Learning*, 94(2):151–167.

Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. 2013. Learning semantic maps from natural language descriptions. Robotics: Science and Systems.

Yezhou Yang, Yiannis Aloimonos, Cornelia Fermuller, and Eren Erdal Aksoy. 2015. Learning the semantics of manipulation action. *arXiv preprint arXiv:1512.01525*.