Every Object Tells a Story

James Pustejovsky
Computer Science Department
Brandeis University
Waltham, MA 02453
jamesp@brandeis.edu

Nikhil Krishnaswamy
Computer Science Department
Brandeis University
Waltham, MA 02453
nkrishna@brandeis.edu

Abstract

Most work within the computational event modeling community has tended to focus on the interpretation and ordering of events that are associated with verbs and event nominals in linguistic expressions. What is often overlooked in the construction of a global interpretation of a narrative is the role contributed by the objects participating in these structures, and the latent events and activities conventionally associated with them. Recently, the analysis of visual images has also enriched the scope of how events can be identified, by anchoring both linguistic expressions and ontological labels to segments, subregions, and properties of images. By semantically grounding event descriptions in their visualizations, the importance of object-based attributes becomes more apparent. In this position paper, we look at the *narrative structure of objects*: that is, how objects reference events through their intrinsic attributes, such as affordances, purposes, and functions. We argue that, not only do objects encode conventionalized events, but that when they are composed within specific habitats, the ensemble can be viewed as modeling coherent event sequences, thereby enriching the global interpretation of the evolving narrative being constructed.

1 Introduction

There has been significant research on the interpretation of events in text, particularly news articles (UzZaman et al., 2013; Pustejovsky et al., 2003; Aguilar et al., 2014). While identifying events and their participants has received much attention in the field, the construction of narratives, stories, scripts, and globally coherent relations between these events, is much more difficult and remains a challenging task (Chambers and Jurafsky, 2009; Rospocher et al., 2016). In this position paper, rather than focus on the semantics associated with event-denoting expressions in language, we discuss the contributions made by object participants in these events, and how these can influence or determine the global narrative event semantics of the text.

The semantic content of events is most often anchored to the matrix predicate of a sentence and the associated event participants, expressed as verbal arguments. Further complicating the interpretation of events is the fact that, while all entities are usually realized as nominal expressions, not all nominals are entities. That is, the participants in events can themselves be events, such as *Heavy rains resulted in flooding*, where both arguments are event-denoting NPs. In such cases, it is clear what role the NPs play in constructing a larger event-based narrative for the text. But there are many ways for a nominal expression to refer to an event, without denoting one. These are called *event-connoting* nominals. Examples include: *agentive nominals*, both occupational and social (dancer, baker, teacher, pilot, neighbor, friend); object *resultative nominals* (debris, mixture, waste, laundry); and all *artifactual nominals* (bread, coffee, desk, house, airplane).

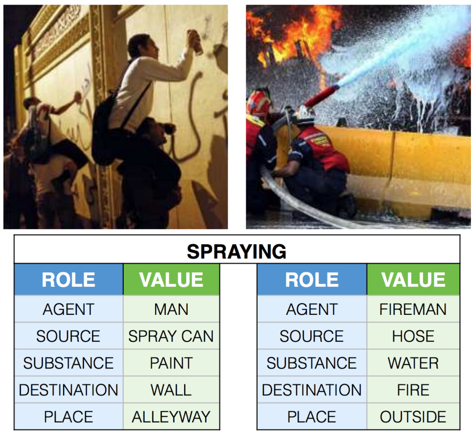
We claim that, while the core structure of a narrative is largely formed through the composition of explicitly mentioned events, that are temporally ordered and constrained through discourse coherence relations, there is another latent narrative structure created from the events and activities associated with object-denoting participants in a text or image.

Proceedings of the Workshop on Events and Stories in the News, pages 1–6
Santa Fe, New Mexico, USA, August 20, 2018.

2 Linguistic Interpretation of Images

The body of work on text and image analysis relies on a number of techniques, e.g., semantic annotation of video; statistical classification for feature detection; heuristic, Markovian, and Bayesian methods for classification of composite events, among many others (Ballan et al., 2011). State of the art includes a variety of metrics to evaluate the robustness of caption generation systems (Anderson et al., 2016), event description (Young et al., 2014), and scene description (Aditya et al., 2015).

Previous work in visual semantic role labeling (e.g., Gupta and Malik (2015); Yatskar et al. (2016)) often involves determining the main activity and participants in an image or scene. In many cases, the activity is closely linked to one of the objects in the scene, and some canonical property of it. For example, in Figure 1 (taken from Yatskar et al. (2016)), we see two examples of a *spraying* event, both closely associated with one particular object in the scene—a spray can or a hose.



SPRAYING			
ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	FIREMAN
SOURCE	SPRAY CAN	SOURCE	HOSE
SUBSTANCE	PAINT	SUBSTANCE	WATER
DESTINATION	WALL	DESTINATION	FIRE
PLACE	ALLEYWAY	PLACE	OUTSIDE

Figure 1: Spraying event with role labels, taken from Yatskar et al. (2016)

Both spray cans and hoses have canonical uses, which is to spray some substance (paint/water); knowing the canonical use of an object allows a human, as a reasoner, to infer what event or subsuming event is being depicted. Similar remarks hold for a tool such as a *chainsaw* (Figure 2a), independent of its role in an explicit *cutting* event (Figure 2b).

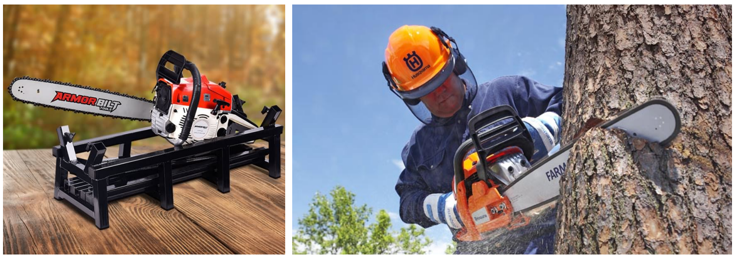


Figure 2: Latent Event (left) and Active Event (right)

In cases where an object is depicted as *violating* its canonical or typical use, the implied narrative becomes yet more interesting. The general "scenario localization" (Pustejovsky, 2013b) and particular types of textual or image narrative connotation can be either encoded or subverted by the presence and depiction/description of objects denoted/depicted within them: subverting the inherent narrative encoded into a particular object introduces a new narrative, vis-à-vis how the object came to be in the situation where is it depicted or described.

Take as an example the events associated with the object denoted by the artifactual nominal *plane*. The prototypical "fly" event can be broken down into the subevents "take off(a)," "translocation(a,b)," "land(b)," (encoded as Generative Lexicon's TELIC role). This forms a canonical *take off-fly-land* narrative associated with a plane. This lexically-encoded narrative can be left uninstantiated (a plane sitting in a hangar) or violated (by a *crash* event).

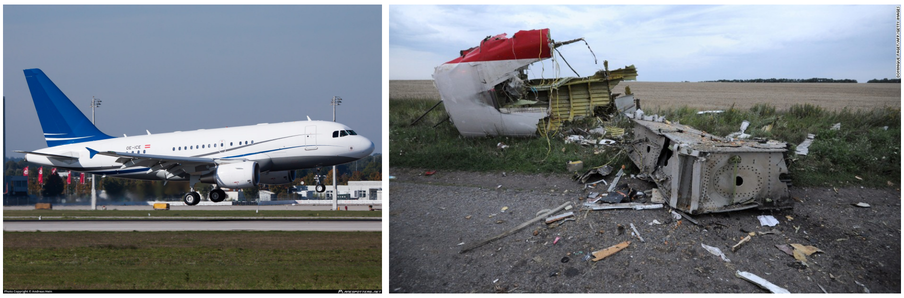


Figure 3: Airplane (left) and airplane debris in a field (right)

The image of debris in a field focuses this interruption or violation of the plane's canonical or typical purpose, as does a hypothetical image caption or snippet of prose narrative (e.g., "People walk among the debris at the crash site of a passenger plane near the village"), that presupposes the existence of the same debris and hence the event *crash* that interrupted the canonical narrative of the *plane*, causing it to enter into the situation where it (and the resultant debris) is described.

3 Habitats and Event-Connoting Expressions

In this section we introduce the specification for how latent event structure is encoded for entity types. Recall that there are three major types of event-connoting nominal expressions: (a) agentive nominals; (b) resultative nominals; and (c) artifactual nominals. Consider first the case of **artifactual nominals**. Following Generative Lexicon (GL) (Pustejovsky, 1995), such nominals are given a feature structure consisting of the word's basic type and its qualia structure. The latent event structure associated with an object is referenced through the qualia: e.g., a food item has a TELIC value of eat; an instrument for writing, a TELIC of $write$; a cup, a TELIC of $hold$, and so forth. Similarly, as mentioned above, the semantics for the noun *plane* carries a TELIC value of fly:

$$(1)\quad \lambda x \exists y \begin{bmatrix} \textbf{plane} & \\ \text{AS} = & \begin{bmatrix} \text{ARG1} = x : e \end{bmatrix} \\ \text{QS} = & \begin{bmatrix} \text{F} = vehicle(x) \\ \text{T} = \lambda z, e[fly(e, z, x)] \end{bmatrix} \end{bmatrix}$$

While convention has allowed us to interpret the entire TELIC expression as modal, this is inadequate for capturing the deeper meaning of functionality, and this introduces the notion of a *habitat*. A habitat can be viewed as the environmental constraints, $\mathcal{C}$, necessary for a latent event to be realized (Pustejovsky, 2013a). Assuming a dynamic semantics for how events are interpreted (Pustejovsky and Moszkowicz, 2011), we can say of an artifact, x, in the appropriate context $\mathcal{C}$, that performing the action π will result in the intended or desired resulting state, $\mathcal{R}$, i.e., $\mathcal{C} \rightarrow [\pi]\mathcal{R}$. That is, if a context $\mathcal{C}$ (a set of contextual factors) is satisfied, then every time the activity of π is performed, the resulting state $\mathcal{R}$ will occur. Hence, while the TELIC event for *plane* is *fly*, it is modal, and the preconditions for such an event must be satisfied, e.g., it has to be oriented properly, have fuel, it is air-worthy, etc., as well as be situated such that it can take off from a source, cruise in a trajectory, and land at a destination. An enriched lexical representation for such information of *plane* would involve far more operational and procedural knowledge than typically associated with the semantics of lexical items, going beyond the normal purview of qualia structure.

For this reason, in order to more richly represent this knowledge structure computationally, we are exploiting the modeling language VoxML (Pustejovsky and Krishnaswamy, 2016; Krishnaswamy and Pustejovsky, 2016), which was initially developed in the context of 3D modeling of language in "multimodal semantic simulations," wherein a computational system can render its interpretation of an event visually, for evaluation or to interact with a human. The VoxML equivalent of the above habitat structure, accounting for placement of the parameters within the embedding space $\mathcal{E}$ is given below:

$$(2)\quad \begin{bmatrix} \textbf{plane} & \\ \text{HABITAT} = & \begin{bmatrix} \text{INTR} = [1] \begin{bmatrix} \text{SRC} = y_1 \in \mathcal{E} \\ \text{DEST} = y_2 \in \mathcal{E} \\ \text{TOP} = top(+Y) \\ \text{DIR} = align(Z, \mathcal{E}_{vec(y_2 - y_1)}) \end{bmatrix} \end{bmatrix} \\ \text{AFFORD\_STR} = & \begin{bmatrix} \text{A}_1 = H[1] \rightarrow [fly(x, y_1, y_2)]\mathcal{R}_{fly}(x) \end{bmatrix} \end{bmatrix}$$

The plane begins from a source heading to a destination. It must remain upright within the medium, and oriented along a trajectory from source to destination. These constraints allow the plane to fulfill its *telic* role “fly,” encoded as an *affordance*, following (Gibson, 1977), and subject to various interpretations (e.g., Steedman (2002); Chao et al. (2015); Osiurak et al.(2017); Poddiakov (2018)).

Now consider the class of **agentive nominals**, such as *pilot*, *pianist*, and the agent from Figure 2b, *chainsaw operator*. All such nouns are typed with a TELIC value referring to a specific event, such as flying a plane, playing piano, or felling trees. Such latent event values are present by virtue of identifying the *typing* of the individual mentioned (in text) or portrayed (in an image).

Finally, consider the class of object **resultative nominals**, such as *debris* (Figure 3b), and the two examples shown below, i.e., *lava flow* and *laundry*.

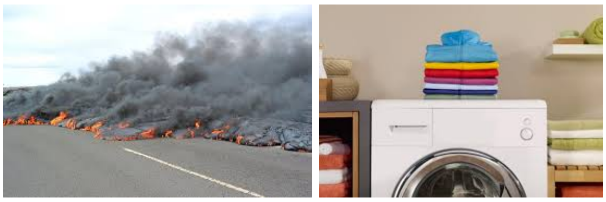


Figure 4: Lava flow (left) and clean laundry (right)

As the name implies, the nominal makes reference to an event bringing about the denotation of the entity (Pustejovsky, 1995; Hovav and Levin, 2010). That is, *debris* is made of parts of a referenced, previously intact object, of which those constitutive parts still exist in some form, but no longer constitute the complete object. *Lava flow*, on the other hand, is actually a polysemous nominal, referring to either an event or the resulting material. Here, the image depicts the rock formation resulting from the event. Finally, the nominal *laundry* has a latent event referencing the resulting state of the clothes (either dirty or clean) being laundered. This image constrains the ambiguity to the clean state, as the folded clothes are situated over the dryer, two additional habitat constraints suggesting this interpretation. This is a signature example of a narrative constructed entirely from the composition of several objects and their associated latent event structures.

4 Future Directions

In this position paper, we argue that there is a potentially rich latent event structure associated with objects in text and images, which can be exploited for enriching the interpretation and construction of narratives. Objects can be seen as encoding latent event structures that, when combined, can create narrative structures of their own. This can potentially provide information for a framework within which to computationally extract linkages between images and text in news stories, and model coherent event sequences, and predict bias or differences of perspective in reporting. In order to test such hypotheses, we are currently annotating and analyzing a number of different corpora to identify both the TELIC roles and affordances associated with objects, as expressed in text and images. These include the Flikr30k (Young et al., 2014; Plummer et al., 2015); the VisualGenome (Krishna et al., 2017); and a subset of the images used for in the Visual Question Answering task from MS COCO (Lin et al., 2014). While there have been some efforts to identify affordances with objects (Chao et al., 2015), it remains a challenging issue to create object-latent event associations at scale. We believe a combination of manual annotation together with automatic extraction techniques for qualia relations (Cimiano and Wenderoth, 2007; Claveau and

Sébillot, 2013) will help in constructing a multimodal lexical resource that reflects the narrative structure of objects.

Acknowledgements

The authors would like to thank the reviewers for their helpful comments. We would also like Tuan Do, Kyeongmin Rim, Marc Verhagen, and David McDonald for discussion on the topic. This work is supported by a contract with the US Defense Advanced Research Projects Agency (DARPA), Contract CwC-W911NF-15-C-0238. Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. 2011. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 602–610. Association for Computational Linguistics.

Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. 2015. Mining semantic affordances of visual object categories. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 4259–4267. IEEE.

Philipp Cimiano and Johanna Wenderoth. 2007. Automatic acquisition of ranked qualia structures from the web. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 888–895.

Vincent Claveau and Pascale Sébillot. 2013. Automatic acquisition of gl resources, using an explanatory, symbolic technique. In *Advances in Generative Lexicon Theory*, pages 431–454. Springer.

James J. Gibson. 1977. The theory of affordances. *Perceiving, Acting, and Knowing: Toward an ecological psychology*, pages 67–82.

S. Gupta and J. Malik. 2015. Visual Semantic Role Labeling. *ArXiv e-prints*, May.

Malka Rappaport Hovav and Beth Levin. 2010. Reflections on manner/result complementarity. In E. Doron, M. Rappaport Hovav, and I. Sichel, editors, *Syntax, Lexical Semantics, and Event Structure*, pages 21–38. Oxford University Press.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Nikhil Krishnaswamy and James Pustejovsky. 2016. Multimodal semantic simulations of linguistically underspecified motion events. In *Spatial Cognition X*, pages 177–197. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

François Osiurak, Yves Rossetti, and Arnaud Badets. 2017. What is an affordance? 40 years later. *Neuroscience & Biobehavioral Reviews*, 77:403–417.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alexander N Poddiakov. 2018. Exploratory and counter-exploratory objects: Design of meta-affordances.

James Pustejovsky and Nikhil Krishnaswamy. 2016. VoxML: A visualization modeling language. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *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, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

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*.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Web Semantics: Science, Services and Agents on the World Wide Web*, 37:132–151.

Mark Steedman. 2002. Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, 25(5-6):723–753.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 1–9.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5534–5542.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.