Multimodal Use of an Upper-Level Event Ontology

Claire Bonial,[1] David Tahmoush,[1] Susan Windisch Brown,[2] and Martha Palmer[2]
[1]U.S. Army Research Lab, Adelphi, Maryland
[2]University of Colorado, Boulder, Boulder, Colorado
Claire.N.Bonial.civ@mail.mil, David.Tahmoush.civ@mail.mil
Susan.Brown@colorado.edu, Martha.Palmer@colorado.edu

Abstract

We describe the ongoing development of a lexically-informed, upper-level event ontology and explore use cases of the ontology. This ontology draws its lexical sense distinctions from VerbNet, FrameNet and the Rich Entities, Relations and Events Project. As a result, the ontology facilitates interoperability and the combination of annotations done for each independent resource. While this ontology is intended to be practical for a variety of applications, here we take the initial steps in determining whether or not the event ontology could be utilized in multimodal applications, specifically to recognize and reason about events in both text and video. We find that the ontology facilitates the generalization of potentially noisy or sparse individual realizations of events into larger categories of events and enables reasoning about event relations and participants, both of which are useful in event recognition and interpretation regardless of modality.

1 Introduction & Background

The valuable computational lexical resources, VerbNet (Kipper et al., 2008), FrameNet (Fillmore et al., 2002), and the Rich Entities, Relations and Events (ERE) annotation project (Song et al., 2015), each provide somewhat distinct information about which eventualities are related syntactically, semantically, or both, and which types of participants are involved in classes of eventualities. VerbNet and FrameNet also involve long-standing and comprehensive annotation efforts, using the class and participant type labels set out in each resource. The resulting annotated corpora have proved to be useful sources of training data for a variety of Natural Language Processing (NLP) systems, including automatic semantic role labeling, word sense disambiguation, and question-answering systems.

Recently, we have also seen an expansion of efforts in both the construction of ontologies as part of the Semantic Web (Berners-Lee, 1998), and research in computer vision (e.g., Fei-Fei & Perona, 2005). These previously disparate threads of research have begun to come together with NLP research in fruitful ways. First, there have been efforts to integrate computational lexical resources into the Semantic Web (e.g., Eckle-Kohler et al., 2014). Progress in this area, however, has been somewhat slow and difficult, given that conversion of resources like FrameNet, which includes quite nuanced and complex ontological relations, into the minimalist Resource Descriptive Framework (RDF) schema used in the Semantic Web is not necessarily a trivial conversion and may involve some loss of information (e.g., Nuzzolese et al., 2011; Scheffczyk et al., 2006). Second, data-driven methods for extracting events from text are increasingly being combined with knowledge-driven methods, such as those based on ontologies, in order to benefit from the strengths of both (Hogenboom et al., 2011). Third, there have been efforts to use both text and visual data jointly to interpret complex scenes (e.g.,

Proceedings of the 4th Workshop on Events: Definition, Detection, Coreference, and Representation, pages 18–26,
San Diego, California, June 17, 2016.

Karpathy et al., 2014). Furthermore, we have seen the promise of integrating ontological information into computer vision research evidenced in the success of ImageNet (Deng et al., 2009), a large-scale ontology of images built upon the structure of WordNet (Fellbaum, 1998).

To further exploit and encourage synergy in all of these research areas, both language processing, vision processing, and the marriage of the two, our goal is to create an upper-level event ontology that provides conceptual coverage for the aforementioned lexical resources, and uses them as the sense inventories that house the linguistic realizations of those concepts in English. This ontology is being developed with multi-modal applications in mind, and in this research we explore the utility of the ontology in a text-processing application as well as a video-processing application. The primary motivations for developing this ontology are two-fold: first, as with most ontologies, the event ontology should allow for reasoning about events and their participants, with a special focus on temporal and causal relations between events. Second, the event ontology should allow for the generalization of specific events, allowing for the detection of similar events in text and facilitating the recognition of the same action being performed by different people under different circumstances in video. The latter aim is especially challenging given that we cannot assume that the level of generalization appropriate for text applications would be the same for video applications.

In this paper, we describe the current structure of the ontology, which is still under development, including a brief description of the resources we are drawing upon for sense distinctions and for temporal and causal relations. The ontology will be openly released to the research community, but is not yet available given that it is still undergoing changes and expansions. Next, we describe preliminary efforts to explore both text and video use cases of the ontology. Pertaining to text, we focus on connecting two different descriptions of the same event via lexical similarity, such as *seize* and *capture*. Next, we examine the feasibility of using ontological relations to improve human activity recognition in videos, focusing on the detection of *pick up and throw* activities, which often occur sequentially.

2 Event Ontology Design

To facilitate compatibility with the Semantic Web, our ontology is being developed using the open-source ontology editor, Protégé (Noy et al., 2000), in OWL format. An early design decision we faced was how to incorporate the lexicons of interest into an ontology. The approach we have chosen is to develop and maintain VerbNet and FrameNet, and the ERE event types as distinct, stand-alone ontologies that are imported into the upper-level event ontology. The individual lexicons and the ontology are linked through the "has_Sense" relation: conceptual nodes in the ontology have senses and associated lexical items spelled out in the lexicons.

2.1 Sense Inventories

Currently, we have successfully implemented both VerbNet and ERE in OWL, since these lexicons have only very shallow hierarchical class structures. However, we are still developing the OWL-implementation of FrameNet, since, as mentioned previously, the ontological structure and inheritance types in FrameNet are quite complex. To this point, we have been developing the FrameNet lexical ontology on an as-needed basis, and including only basic inheritance links within the FrameNet ontology. However, we are exploring the feasibility of adopting an existing OWL-implementation of FrameNet (e.g., Scheffczyk et al., 2006), and importing this directly into the upper-level ontology. Brief descriptions of each of the lexical resources included in the ontology are given below, with explanations of how the sense distinctions vary across each resource.

VerbNet, based on the work of Levin (1993), groups verbs into "classes" based on compatibility with certain "diathesis alternations" or syntactic alternations (e.g., *She loaded the wagon with hay* vs. *She loaded hay into the wagon*). Although the groupings are primarily syntactic, the classes do share semantic features as well, since, as Levin posited, the syntactic behavior of a verb is largely determined by its meaning.

VerbNet includes a semantic representation for each usage example demonstrating a characteristic diathesis alternation of a class. For example, in the Throw class, the following alternate is listed with its

semantic representation, based on individual semantic predicates (e.g., **motion, cause, exert_force**):

Example: "Steve tossed the ball to the garden."
Roles: Agent Verb Theme Destination
Semantic Predicates
EXERT_FORCE(during(E0), Agent, Theme)
CONTACT(during(E0), Agent, Theme)
MOTION(during(E1), Theme)
not(CONTACT(during(E1), Agent, Theme))
not(LOCATION(start(E1), Theme, Destination))
LOCATION(end(E1), Theme, Destination)
CAUSE(Agent, E1)
meets(E0, E1)

This representation is intended to break the event down into smaller semantic elements, given as the predicates. The predicates are organized with respect to the time of the event (designated as 'E'); thus they can apply *during*, at the *start* or at the *end* of an event. Although somewhat complex for this event, the above representation is meant to capture the fact that Steve (Agent) is in contact with and exerts force on a ball (Theme); he then releases (is not in contact with) the ball and the ball is in motion; the ball's location at the end of the motion event is at the garden (Destination), where it was not located at the start of the event; Steve causes this event as the Agent having thrown the ball. Notice that although this semantic interpretation captures many of the salient semantic components of a throwing event, it may not capture the salient *visual* aspects of a throwing event, which may stem more from the sequential motions of the body and body parts.

Within the ontology, VerbNet will serve as a lexicon imported into the ontology, and it will therefore provide one set of sense distinctions for the English lexical items that denote concepts within the ontology. Because class membership in VerbNet is in part based on syntactic information, VerbNet captures the level of sense distinctions that are clearly evidenced by differences in syntactic behaviors.

FrameNet, based on Fillmore's frame semantics (Fillmore, 1976; Fillmore & Baker, 2001), groups verbs, nouns and adjectives into "frames" based on words or "frame elements" that evoke the same semantic frame: a description of a type of event, relation, or entity and the participants in it. For example, the Apply_heat frame includes the frame elements Cook, Food, Heating_instrument, Temperature_setting, etc. The "net" of frames makes up a rather complex ontological network, including simple "is_A" inheritance relations as well as more complex relations such as Precedes and Perspective_on. FrameNet will serve as another lexicon within the ontology, providing a different set of sense distinctions for the lexical items denoting concepts. Since the classification of FrameNet is purely semantic and based on shared frame elements, the sense distinctions and distinctions between participant types made in FrameNet are often more fine-grained than VerbNet. For example, given the sentence *Sally fried an egg*, VerbNet would label *Sally* with the traditional semantic role label Agent, while FrameNet would label *Sally* with the more semantically specified label of Cook.

The Rich Entities, Relations, and Events (ERE) project is based on the Automatic Content Extraction (ACE) project's semantic role annotation schema (Doddington et al., 2004). The goal of the ERE project is to mark up the events (and other types of relations; i.e. "eventualities") and the entities involved in them, and to mark coreference between these. This provides a somewhat shallow representation of the meaning of the text. The ERE schema will also serve as a sort of lexicon imported into the ontology, with its event type and subtype designations serving as links to the lexical items marked up with that designation. ERE annotated eventualities are limited to certain types and subtypes of special interest within the defense community, with top-level types referred to as *Life, Movement, Transaction, Business, Conflict, Manufacture, Contact, Personnel* and *Justice* events. Thus, the sense distinctions made by this resource are grounded in practical considerations of what event types are deemed to be of interest, and therefore offer very different insights and information into related events when compared with either FrameNet or VerbNet.

2.2 Ontology Structure & Relations

There are many critical decisions to be made when determining what concepts and relations to represent in an ontology; thus, we draw upon the decisions made in the established ontologies WordNet

(Fellbaum, 1998), the Suggested Upper Merged Ontology, SUMO (Pease, 2002), the Descriptive Ontology for Linguistic and Cognitive Engineering, DOLCE (Masolo et al., 2003), and The Basic Formal Ontology, BFO (Smith & Grenon, 2002). Similar to the structures of each of these existing ontologies, we have selected a top level concept, Entity, with an initial distinction between Endurant and Perdurant entities. Borrowing heavily from DOLCE, we have developed the following definitions. We define "Entity" as *a unique object or set of objects in the world – for instance a specific person, place or organization – that typically functions as a participant*. We define "Endurants" as *those entities that can be observed/perceived as a complete concept, no matter which given snapshot of time – were we to freeze time, we would still be able to perceive the entire endurant*. We define "Perdurants" as *those entities for which only a part exists if we look at them at any given snapshot in time. Variously called events, processes phenomena, or activities and states, perdurants have temporal parts or spatial parts and participants*.

We also need to capture richer information about the temporal and causal relations between events than any of the lexical resources described thus far are currently capturing independently. To ensure that the ontology captures temporal and causal relations of utility within NLP, we use relations from the established Richer Event Description (RED) project (Ikuta et al., 2014). Like ERE, the RED project also aims to markup text with mentions of eventualities and entities, but the primary focus of RED is to represent the temporal and causal relationships between those eventualities. The final goal is to produce annotations rich enough that a computer, using complex inferencing, co-reference, and domain-specific algorithms, would be able to construct an accurate timeline of when the events in a given document occur relative to any fixed dates present and relative to one another (e.g., automatically constructed timelines of medical histories). RED builds on THYME (Styler et al., 2014), a temporal relationship annotation of clinical data that is based on TimeML (Pustejovsky et al., 2010). The temporal relations are quite fine-grained, including *Before, Before+overlap, Overlap, and CONTAIN*. These labels are further distinguished with causal labels where appropriate: *Before/Causes, Before/Preconditions*. To anchor the events into a timeline, RED links the event to a document time or section time where applicable, and marks up explicit references to time in the document.

2.3 Snapshot of the Ontology

The construction of the ontology is still underway, and has involved a combination of bottom-up and top-down approaches. As ERE event types provide useful constraints on which events to focus on initially, our efforts generally begin with an examination of a particular ERE event type, a comparison of sense distinctions and associated lexical items made in VerbNet and FrameNet, followed by a preliminary fleshing-out of one area of the ontology. At this point, we have situated the top level ERE event types *Life*, *Conflict*, *Contact*, *Justice* and *Personnel*. Thus, we have also situated most of the subtypes within these event types, although our approach involves some iterative refinement of the ontology's class structure.

A simple example portion of the ontology, with sense mappings to VerbNet only, is shown in Figure 1. Within each of the lexical classes, the individual lexical items denoting senses of a concept are listed within each resource. For example, the VerbNet Birth class lists the lexical items *bear, birth, deliver, father, mother, sire, spawn*, etc. The parallel FrameNet Being_born frame includes just *born* and the phrase *come into the world*. ERE realizations include any lexical item tagged as a trigger indicating this type of event during annotation, such as the verb *born* and the noun *birth*.

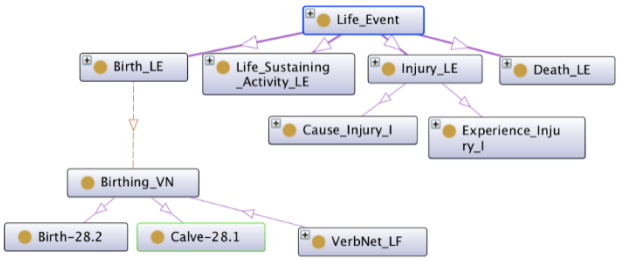

Figure 1: Life_Event extract of event ontology.

Events within the ontology are also related via temporal and causal relations, such as "has_Result" and "has_Precondition." Here, for example, Birth life events are linked to the Life state, Alive (a daughter node to Stative_Perdurant), through the "has_Result" link: once something is born, it is alive. While

some of these relations have been developed (and are still under development) specifically for the upper-level ontology, other relations stem directly from the RED project.

3 Use Cases

As mentioned above, the desire to robustly identify and corefer events, either in text or in video, is a primary motivating factor for our development of the ontology. We also wish to be able to reason about the events detected, and draw plausible inferences. Both of these goals require a level of abstraction that can facilitate the detection of two different instances of the same event, whether they are textual or visual. For instance, textual event descriptions may use quite different language, such as *Stock prices rose precipitously*, and *The Stock Market leapt ahead.* Visual events might have different types of agents, performing the same action in different ways. There is preliminary evidence that VerbNet argument structure and semantic predicates can provide a useful level of abstraction, that can serve both language processing and vision processing aims. Prior work demonstrated VerbNet's effectiveness in providing the foundation for Parameterized Action Representations (PARs), a framework for translating verbal instructions into processing commands for a virtual agent (Badler et al., 1999; Badler et al, 2000). The thematic roles indicate the participants in the action, and the semantic predicates help to define the planning goals the virtual agent needs to properly execute the desired action. The level of abstraction encoded in PARs generalized surprisingly well to different sizes and types of agents, simplifying the task of using motion capture to expand the coverage of actions the virual agents could perform (Bindiganavale et al., 2000). The PAR application focuses on the generation of virtual action videos rather than analysis, but it provides encouraging evidence that the same level of abstraction could benefit analysis as well.

In the sections to follow, we introduce the ways in which our ontology and its connections to established lexical resources could be uniquely valuable for both a text and video processing application.

3.1 Text Use Case

For supervised machine learning systems, instances unseen in training data are problematic. Access to related terms found in lexical resources can allow them to generalize training data to additional instances.

Liu et al. (2014) identify problems even with human annotation of events, noting that in the EventCorefBank corpus (Bejan and Harabajiu, 2010), "seizing 12 Somali and 11 Yemeni nationals" and "capturing 23 of the raiders" had not been identified as the same event. Our ontology, in which the lexical items *seize* and *capture* are linked to the same class, could be helpful in automatically connecting these two mentions as the same event. In the discussion of their system of event coreference, Liu et al. also noted that "[event coreference] can possibly be improved by other types of event relations, such as subevent relations." Such relations are an integral part of our event ontology, along with causal, precondition and postcondition relations.

The lexical resources themselves can also be limited in their coverage. For example, consider again the Life_Event portion of the ontology shown in Figure 1. FrameNet lists two entries, *born* and *come into the world* in a frame that would be associated with Birth_LifeEvent. This, too, can lead to data sparsity, but the ontology facilitates pinpointing other lexical items that are more generally related to the entries in FrameNet's semantically fine-grained Being_born frame. Thus, the ontology facilitates some interoperability between the individual lexical resources and makes explicit some of the previously unseen relations between them.

3.2 Video Use Case

Another potential application for an event ontology is for action recognition in video. Human activity recognition has already been explored using images and video. Activity recognition techniques can be grouped into data-driven (Ye et al., 2012) and knowledge-driven (Chen et al., 2012a) approaches. Data-driven techniques use machine learning approaches to discern an activity from the training data. Space-time methods such as space-time volumes, spatio-temporal features, and trajectories have been successful. For classification, generic approaches like support vector machines and hidden Markov models have performed well.

So far, most contestants at a recent Action Classification Challenge at the 2013 International Conference on Computer Vision (ICCV) have utilized low-level features over higher-level class attributes and ontologies because they traditionally have been more effective. Nonetheless, an action ontology can provide a description of the activity using well-structured terminology with a number of properties that are measureable. A well-built ontology could be used, understood, and shared between humans and computers (Gu et al., 2004; Riboni et al., 2009; Chen et al., 2012b) and, as with text applications can facilitate the generalization of specific instances of actions to enable more consistent detection of the same activity, done by different people in different contexts. Furthermore, an action ontology can provide information about the temporal and causal relationships between component actions involved in an activity, potentially improving recognition in ambiguous cases.

In related work, an inventory of attributes of human activities (largely activities related to sports and playing musical instruments) has been developed with a focus on attributes that are visually salient (Jiang et al., 2013). The attributes listed in this resource are quite simple, primarily organized around what body parts are used. For a given action, common attributes of certain body parts or areas of the body are listed for each action (e.g., for *throwing*, Body Part Articulation-Arm = One Arm Raised Head Level).

To a limited extent, these attributes provide a break-down of some of the component parts of an action by detailing some of the ways in which individual body parts move during an activity. However, these attributes are not ordered with respect to time, and many attributes apply throughout an activity instead of expressing finer-grained motions within an activity (e.g., Outdoor). Thus, although this inventory captures many of the aspects of an action that are visually salient, it fails to capture both temporal information and the event-semantics information that is salient in text, which VerbNet's semantic predicates capture.

Therefore, we see an opportunity to explore combining some of these visually salient action attributes with the semantic components of events laid out in VerbNet, and integrating these at the lowest levels of the ontology so that they are clearly temporally and causally related. For example, *throw* events can be broken down into the following time-ordered attributes based on VerbNet semantic predicates (described in Section 2.1):

- Time1: Agent physically grasps Theme – **Contact**(Agent, Theme)
- Time2: Exert force propelling Theme from Agent – **Exert_Force**(Agent, Theme)
- Time3: Release theme – ***not***(**Contact**(Agent, Theme))
- Time4: Theme is in motion – **motion**(Theme)

Situated in the ontology, these attributes would be related via RED temporal/causal relations, making explicit the fact that **Contact** is in a *Before+Overlapping/Preconditions* relation with **Exert_Force**, and that **Exert_Force** is in a *Before/Causes* relation with **Motion.** Admittedly, to break down all events into these smaller components would be a massive undertaking; thus, we would consider only a subset of sufficiently physical actions, and we begin here by simply exploring whether or not breaking down a single complex activity, *pick up and throw*, into its components seems to aid in recognition.

3.2.1 Approach: Skeletonized Video

As mentioned previously, human activity recognition has been explored using images and video, and one benchmark is the MSR-Action3D dataset (Jiang et al., 2013). It includes 20 actions performed by 10 subjects, and each subject performs an action 3 times. An example is shown in Figure 2 for *high wave* where the motion of the arms, legs, head, and torso are shown with the depth dimension removed.

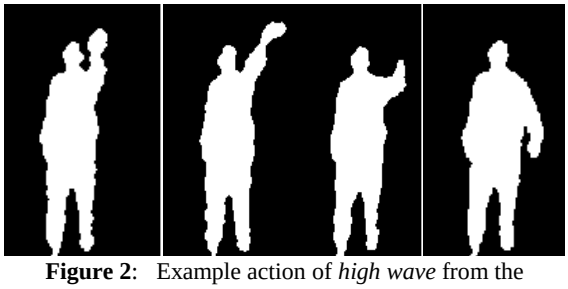

Figure 2: Example action of *high wave* from the MSR-Action3D dataset.

For this work, 20 joint positions are tracked using a skeleton tracker and compiled into a time series of joints *i* depicted as $\boldsymbol{p}i(t) = (xi(t); yi(t); zi(t))$ at a frame *t*. The coordinates are then normalized to re-

duce dependencies on height, initial body orientation and location. An example of a skeletonized motion is shown in Figure 3.

3.2.2 Preliminary Exploration: Pick Up, Throw

Preliminary work has shown the capability to recognize activities when they are isolated (Tahmoush, 2015). This is similar to recognizing a single word on a piece of paper. However, in longer videos, it is often necessary to segment the video into pieces when there are multiple activities that occur.

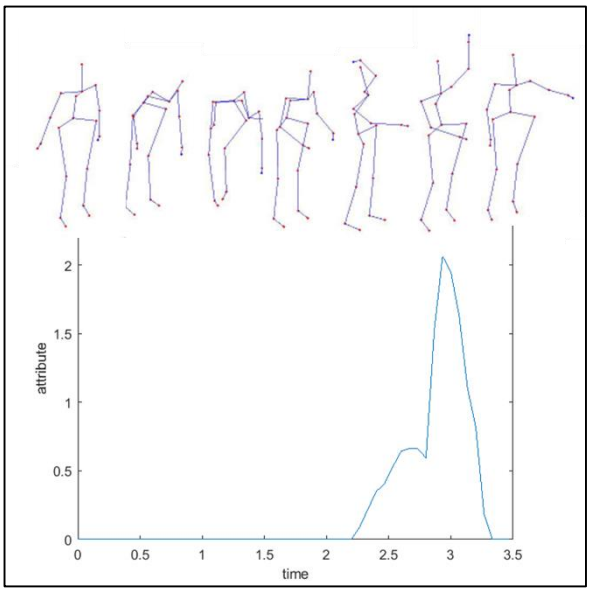

Figure 3: Frames from a skeletonized motion for *pick up and throw* with the measured attribute Body Part Articulation-Arm = One Arm Raised Head Level. Note that the attribute segments out the time period for the *throw* portion.

We look at the compound action of *pick up and throw*, which is part of the MSR-Action3D dataset, which also includes *high throw* and *bend*. When combined, the latter create an activity very similar to *pick up and throw*. In fact, a primary source of misclassifications on this dataset is *pick up and throw* being mistaken for *bend*, as shown in Figure 4.

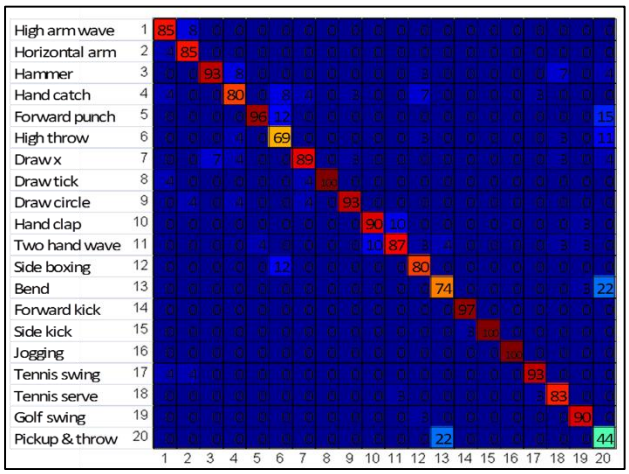

Figure 4: Confusion matrix for activities without using an ontology.

Results of our initial research recognizing smaller activities in larger datasets has shown that motion attributes can perform the important segmentation of videos which can enable improved recognition. In the case of *pick up and throw*, the *throw* activity can be segmented from the *pick up* activity using just one attribute, the Body Part Articulation-Arm = One Arm Raised Head Level attribute that is listed for *throw* (within the previously described Jiang et al. (2013) inventory of attributes). Thus, the use of fast, measurable attributes from an action ontology can effectively segment the video into smaller, more easily recognizable activities without running a costly suite of classifiers. This is illustrated in Figure 3 along with the skeletonized motion.

Our research shows that using the attribute Body Part Articulation-Arm = One Arm Raised Head Level effectively segments the data in 23 out of the 27 *pick up and throw* cases. Visual inspection on the 4 failing cases showed skeletonization failure – the skeletal joints could not be extracted from the video data, thus no skelotonization was produced. Additional visual inspection showed successful segmentation on three cases of *sidearm throwing* and one case of an *underhand toss*, all aberrant cases for *throw* in this data collection. Depending upon how narrowly you define a *throw* action, these could be viewed as false positives. However, this implies both that the attributes allow for some generalization in recognizing *throwing* of all types, and that the attributes may perform well in the case of large variability in the interpretation of the activity.

4 Conclusions & Future Work

An event ontology that could be equally relevant to text processing and language processing would be especially useful for multi-modal processing, and would allow the same generalizations and inferences to be drawn whether the input was textual or visual or both. It is too soon to know if the event ontology described here will achieve that lofty goal, but the possibility is real. In general, we see the potential for ontologies to improve human activity recognition by allowing for more complex actions to be broken down into more easily recognizable activities, as well as easily identifiable, defining attributes of those activities. Therefore, as we further develop the ontology, we are exploring situating fine-grained action attributes within the ontology. Unlike the existing action inventory referenced, this will provide more sophisticated ontological relations, include temporal and causal relations between attributes. We will also continue to explore compatibility with and/or integration of existing ontologies, including Cyc (Reed & Lenat, 2002) and the Emotion Ontology (Hastings et al., 2011).

Acknowledgments

We gratefully acknowledge the support of DARPA DEFT - FA-8750-13-2-0045 and DTRA HDTRA1 -16-1-0002/Project # 1553695, eTASC - Empirical Evidence for a Theoretical Approach to Semantic Components. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or the US government.

References

Badler, Norm, Martha Palmer and Rama Bindiganavale. (1999). Animation Control for Real-Time Virtual Humans. *Communications of the ACM. 42(8):65-73*.

Badler, Norm, Rama Bindiganavale, Jan Allbeck, William Schuler, Liwei Zhao, and Martha Palmer. (2000). A Parameterized Action Representation for Virtual Human Agents. *Embodied Conversational Agents.* MIT Press, pgs. 256-284.

Berners-Lee, T. (1998). Semantic web road map.

Bejan, C., & Harabagiu, S. (2010). Unsupervised event coreference resolution with rich linguistic features. Proceedings of the 48th Annual Meeting of the Association for the Association for Computational Linguistics(July), 1412–1422.

Bindiganavale, Rama, William Schuler, Jan Allbeck, Norm Badler, Aravind Joshi,and Martha Palmer. (2000). Dynamically altering agent behaviors using natural language instructions. *Proceedings of Autonomous Agents 2000.*

Bindiganavale, Rama, (2000). *Building parameterized action representations from observation*, Dissertation, University of Pennsylvania.

Chen, L., Hoey, J., Nugent, C. D., Cook, D. J., & Yu, Z. (2012a). Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 42(6), 790-808.

Chen, L., Nugent, C. D., & Wang, H. (2012b). A knowledge-driven approach to activity recognition in smart homes. *IEEE Transactions on Knowledge and Data Engineering, 24*(6), 961-974.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009, June. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 248-255). IEEE.

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., & Weischedel, R. M. (2004, May). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *LREC* (Vol. 2, p. 1).

Eckle-Kohler, J., McCrae, J., & Chiarcos, C. (2014). lemonUby-a large, interlinked, syntactically-rich resource for ontologies. *Semantic Web Journal, submitted. special issue on Multilingual Linked Open Data*.

Fei-Fei, L. and Perona, P., 2005, June. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 2, pp. 524-531). IEEE.

Fellbaum, C. (Ed.) (1998.) *WordNet: An Electronic Lexical Database.* MIT Press.

Fillmore, C. J. (1976). Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, *280*(1), 20-32.

Fillmore, C. J., & Baker, C. F. (2001, June). Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*.

Fillmore, Charles J., Christopher R. Johnson, and Miriam R.L. Petruck. (2002.) Background to FrameNet. *International Journal of Lexicography*, 16(3):235-250.

Gu, T., Wang, X. H., Pung, H. K., & Zhang, D. Q. (2004). An ontology-based context model in intelligent environments. In *Communication networks and distributed systems modeling and simulation conference*, 270-275.

Hastings, J., Ceusters, W., Smith, B., & Mulligan, K. (2011). The emotion ontology: enabling interdisciplinary research in the affective sciences. In*Modeling and Using Context* (pp. 119-123). Springer Berlin Heidelberg.

Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F.An overview of event extraction from text. Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference (ISWC 2011).

Ikuta, R., Styler IV, W. F., Hamang, M., O'Gorman, T., & Palmer, M. (2014). Challenges of Adding Causation to Richer Event Descriptions. *ACL 2014*, 12.

Jiang, Y.G., Liu, J., Zamir, A.R., Laptev, I., Piccardi, M., Shah, M., & Sukthankar. (2013). R.: THUMOS challenge: action recognition with a large number of classes. http://crcv.ucf.edu/ICCV13-Action-Workshop/.

Karpathy, A., Joulin, A. and Li, F.F.F., 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems* (pp. 1889-1897).

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. (2008.) A large-scale classification of English verbs. *Language Resources and Evaluation Journal*, 42: pp. 21– 40.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Li, W., Zhang, Z., & Liu, Z. (2010). Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW),* 9-14.

Liu, Z., J. Araki, E.H. Hovy, and T. Mitamura. (2014). Supervised Within-Document Event Coreference using Information Propagation. *Proceedings of the LREC conference*. Reykjavik, Iceland.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., & Schneider, L. (2003). Dolce: a descriptive ontology for linguistic and cognitive engineering. *WonderWeb Project, Deliverable D*, *17*.

Noy, N. F., Sintek, M., Decker, S., Crubézy, M., Fergerson, R. W., & Musen, M. A. (2001). Creating semantic web contents with protege-2000.*IEEE intelligent systems*, (2), 60-71.

Nuzzolese, A. G., Gangemi, A., & Presutti, V. (2011, June). Gathering lexical linked data and knowledge patterns from FrameNet. In *Proceedings of the sixth international conference on Knowledge capture* (pp. 41-48). ACM.

Pease, A., Niles, I., & Li, J. (2002, July). The suggested upper merged ontology: A large ontology for the semantic web and its applications. In*Working notes of the AAAI-2002 workshop on ontologies and the semantic web* (Vol. 28).

Pustejovsky, J., Lee, K., Bunt, H. and Romary, L., (2010). ISO-TimeML: An International Standard for Semantic Annotation. In *LREC*.

Reed, S. L., & Lenat, D. B. (2002, July). Mapping ontologies into Cyc. In*AAAI 2002 Conference Workshop on Ontologies For The Semantic Web* (pp. 1-6).

Riboni, D., & Bettini, C. (2009). Context-aware activity recognition through a combination of ontological and statistical reasoning. In *Ubiquitous Intelligence and Computing*, 39-53.

Scheffczyk, J., Baker, C. F., & Narayanan, S. (2006). Ontology-based reasoning about lexical resources. In *Proc. of OntoLex* (pp. 1-8).

Smith, B., & Grenon, P. (2002). Basic formal ontology. *Draft. Downloadable at http://ontology. buffalo. edu/bfo*.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., ... & Ma, X. (2015, June). From light to rich ERE: annotation of entities, relations, and events. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT*(pp. 89-98.

Styler, William, F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova and James Pustejovsky. (2014). Temporal Annotation in the Clinical Domain. *Transactions of the Association of Computational Linguistics*, 2, pp. 143-154.

Tahmoush, D. (2015). Applying Action Attribute Class Validation to Improve Human Activity Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 15-21).

Ye, J., Dobson, S., & McKeever, S. (2012). Situation identification techniques in pervasive computing: A review. *Pervasive and mobile computing*, *8*(1), 36-66.