Grounded Semantic Role Labeling

Shaohua Yang[1], **Qiaozi Gao**[1], **Changsong Liu**[1], **Caiming Xiong**[2],
Song-Chun Zhu[3], and **Joyce Y. Chai**[1]

[1]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824
[2]Metamind, Palo Alto, CA 94301
[3]Center for Vision, Cognition, Learning, and Autonomy, University of California, Los Angeles, CA 90095
{yangshao,gaoqiaoz,cliu,jchai}@cse.msu.edu
cmxiong@metamind.io, sczhu@stat.ucla.edu

Abstract

Semantic Role Labeling (SRL) captures semantic roles (or participants) such as `agent`, `patient`, and `theme` associated with verbs from the text. While it provides important intermediate semantic representations for many traditional NLP tasks (such as information extraction and question answering), it does not capture grounded semantics so that an artificial agent can reason, learn, and perform the actions with respect to the physical environment. To address this problem, this paper extends traditional SRL to grounded SRL where arguments of verbs are grounded to participants of actions in the physical world. By integrating language and vision processing through joint inference, our approach not only grounds explicit roles, but also grounds implicit roles that are not explicitly mentioned in language descriptions. This paper describes our empirical results and discusses challenges and future directions.

1 Introduction

Linguistic studies capture semantics of verbs by their frames of thematic roles (also referred to as semantic roles or verb arguments) (Levin, 1993). For example, a verb can be characterized by `agent` (i.e., the animator of the action) and `patient` (i.e., the object on which the action is acted upon), and other roles such as `instrument`, `source`, `destination`, etc. Given a verb frame, the goal of Semantic Role Labeling (SRL) is to identify linguistic entities from the text that serve different thematic roles (Palmer et al., 2005; Gildea and Jurafsky, 2002; Collobert et al., 2011; Zhou and Xu, 2015). For example, given the sentence *the woman takes out a cucumber from the refrigerator.*, *takes out* is the main verb (also called `predicate`); the noun phrase *the woman* is the `agent` of this action; *a cucumber* is the `patient`; and *the refrigerator* is the `source`.

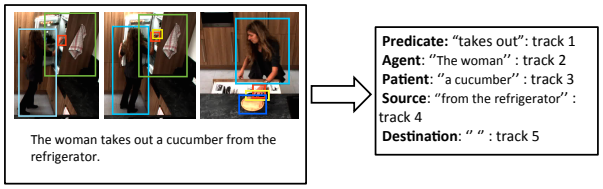

Figure 1: An example of grounded semantic role labeling for the sentence *the woman takes out a cucumber from the refrigerator*. The left hand side shows three frames of a video clip with the corresponding language description. The objects in the bounding boxes are tracked and each track has a unique identifier. The right hand side shows the grounding results where each role including the implicit role (`destination`) is grounded to a track id.

SRL captures important semantic representations for actions associated with verbs, which have shown beneficial for a variety of applications such as information extraction (Emanuele et al., 2013) and question answering (Shen and Lapata, 2007). However, the traditional SRL is not targeted to represent verb semantics that are grounded to the physical world so that artificial agents can truly understand the ongoing activities and (learn to) perform the specified actions. To address this issue, we propose a new task on grounded semantic role labeling.

Figure 1 shows an example of grounded SRL.

Proceedings of NAACL-HLT 2016, pages 149–159,
San Diego, California, June 12-17, 2016.

The sentence *the woman takes out a cucumber from the refrigerator* describes an activity in a visual scene. The semantic role representation from linguistic processing (including implicit roles such as `destination`) is first extracted and then grounded to tracks of visual entities as shown in the video. For example, the verb phrase *take out* is grounded to a trajectory of the right hand. The role `agent` is grounded to the person who actually does the *take-out* action in the visual scene (*track 1*) ; the `patient` is grounded to the cucumber taken out (*track 3*); and the `source` is grounded to the refrigerator (*track 4*). The implicit role of `destination` (which is not explicitly mentioned in the language description) is grounded to the cutting board (*track 5*).

To tackle this problem, we have developed an approach to jointly process language and vision by incorporating semantic role information. In particular, we use a benchmark dataset (TACoS) which consists of parallel video and language descriptions in a complex cooking domain (Regneri et al., 2013) in our investigation. We have further annotated several layers of information for developing and evaluating grounded semantic role labeling algorithms. Compared to previous works on language grounding (Tellex et al., 2011; Yu and Siskind, 2013; Krishnamurthy and Kollar, 2013), our work presents several contributions. First, beyond arguments explicitly mentioned in language descriptions, our work simultaneously grounds explicit and implicit roles with an attempt to better connect verb semantics with actions from the underlying physical world. By incorporating semantic role information, our approach has led to better grounding performance. Second, most previous works only focused on a small number of verbs with limited activities. We base our investigation on a wider range of verbs and in a much more complex domain where object recognition and tracking are notably more difficult. Third, our work results in additional layers of annotation to part of the TACoS dataset. This annotation captures the structure of actions informed by semantic roles from the video. The annotated data is available for download [1]. It will provide a benchmark for future work on grounded SRL.

[1] `http://lair.cse.msu.edu/gsrl.html`

2 Related Work

Recent years have witnessed an increasing amount of work in integrating language and vision, from earlier image annotation (Ramanathan et al., 2013; Kazemzadeh et al., 2014) to recent image/video caption generation (Kuznetsova et al., 2013; Venugopalan et al., 2015; Ortiz et al., ; Elliott and de Vries, 2015; Devlin et al., 2015), video sentence alignment (Naim et al., 2015; Malmaud et al., 2015), scene generation (Chang et al., 2015), and multimodel embedding incorporating language and vision (Bruni et al., 2014; Lazaridou et al., 2015).

What is more relevant to our work here is recent progress on grounded language understanding, which involves learning meanings of words through connections to machine perception (Roy, 2005) and grounding language expressions to the shared visual world, for example, to visual objects (Liu et al., 2012; Liu and Chai, 2015), to physical landmarks (Tellex et al., 2011; Tellex et al., 2014), and to perceived actions or activities (Tellex et al., 2014; Artzi and Zettlemoyer, 2013).

Different approaches and emphases have been explored. For example, linear programming has been applied to mediate perceptual differences between humans and robots for referential grounding (Liu and Chai, 2015). Approaches to semantic parsing have been applied to ground language to internal world representations (Chen and Mooney, 2008; Artzi and Zettlemoyer, 2013). Logical Semantics with Perception (LSP) (Krishnamurthy and Kollar, 2013) was applied to ground natural language queries to visual referents through jointly parsing natural language (combinatory categorical grammar (CCG)) and visual attribute classification. Graphical models have been applied to word grounding. For example, a generative model was applied to integrate And-Or-Graph representations of language and vision for joint parsing (Tu et al., 2014). A Factorial Hidden Markov Model (FHMM) was applied to learn the meaning of nouns, verbs, prepositions, adjectives and adverbs from short video clips paired with sentences (Yu and Siskind, 2013). Discriminative models have also been applied to ground human commands or instructions to perceived visual entities, mostly for robotic applications (Tellex et al., 2011; Tellex et al., 2014). More recently, deep learn-

ing has been applied to ground phrases to image regions (Karpathy and Fei-Fei, 2015).

3 Method

We first describe our problem formulation and then provide details on the learning and inference algorithms.

3.1 Problem Formulation

Given a sentence S and its corresponding video clip V, our goal is to ground explicit/implicit roles associated with a verb in S to video tracks in V. In this paper, we focus on the following set of semantic roles: {`predicate, patient, location, source, destination, tool`}. In the cooking domain, as actions always involve hands, the `predicate` is grounded to the hand pose represented by a trajectory of relevant hand(s). Normally `agent` would be grounded to the person who does the action. As there is only one person in the scene, we thus ignore the grounding of the `agent` in this work.

Video tracks capture tracks of objects (including hands) and locations. For example, in Figure 1, there are 5 tracks: human, hand, cucumber, refrigerator and cutting board. Regarding the representation of locations, instead of discretization of a whole image to many small regions(large search space), we create locations corresponding to five spatial relations (center, up, down, left, right) with respect to each object track, which means we have 5 times number of locations compared with number of objects. For instance, in Figure 1, the `source` is grounded to the center of the bounding boxes of the refrigerator track; and the `destination` is grounded to the center of the cutting board track.

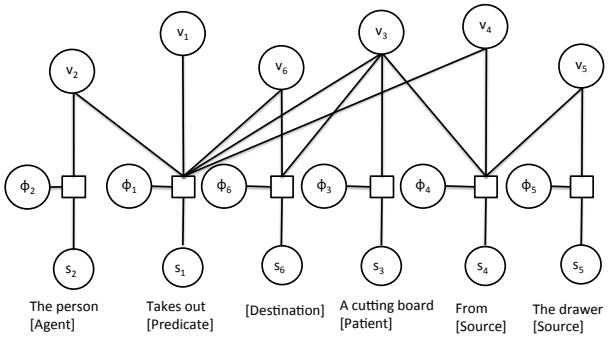

Figure 2: The CRF structure of sentence "*the person takes out a cutting board from the drawer*". The text in the square bracket indicates the corresponding semantic role.

We use Conditional Random Field(CRF) to model this problem. An example CRF factor graph is shown in Figure 2. The CRF structure is created based on information extracted from language. More Specifically, $s_1, ..., s_6$ refers to the observed text and its semantic role. Notice that s_6 is an implicit role as there is no text from the sentence describing `destination`. Also note that the whole prepositional phrase "from the drawer" is identified as the `source` rather than "the drawer" alone. This is because the prepositions play an important role in specifying location information. For example, "near the cutting boarding" is describing a location that is near to, but not exactly at the location of the cutting board. Here $v_1, ..., v_6$ are grounding random variables which take values from object tracks and locations in the video clip, and $\phi_1, ..., \phi_6$ are binary random variables which take values {0,1}. When ϕ_i equals to 1, it means v_i is the correct grounding of corresponding linguistic semantic role, otherwise it is not. The introduction of random variables ϕ_i follows previous work from Tellex and colleagues (Tellex et al., 2011), which makes CRF learning more tractable.

3.2 Learning and Inference

In the CRF model, we do not directly model the objective function as:

$$p(v_1, ..., v_k|S, V)$$

where S refers to the sentence, V refers to the corresponding video clip and v_i refers to the grounding variable. Because the gradient based learning method needs the expectation of $v_1, ..., v_k$, which is infeasible, we instead use the following objective function:

$$P(\phi|s_1, s_2, \ldots, s_k, v_1, v_2, \ldots, v_k, V)$$

where ϕ is a binary random vector $[\phi_1, ..., \phi_k]$, indicating whether the grounding is correct. In this way, the objective function factorizes according to the structure of language with local normalization at each factor.

Gradient ascent with L2 regularization was used for parameter learning to maximize the objective function:

$$\frac{\partial L}{\partial w} = \sum_i F(\phi_i, s_i, v_i, V) - \sum_i \mathbb{E}_{P(\phi_i|s_i,v_i,V)} F(\phi_i, s_i, v_i, V)$$

where F refers to the feature function. During the training, we also use random grounding as negative samples for discriminative training.

During inference, the search space can be very large when the number of objects in the world increases. To address this problem we apply beam search to first ground roles including `patient`, `tool`, and then other roles including `location`, `source`, `destination` and `predicate`.

4 Evaluation

4.1 Dataset

We conducted our investigation based on a subset of the TACoS corpus (Regneri et al., 2013). This dataset contains a set of video clips paired with natural language descriptions related to several cooking tasks. The natural language descriptions were collected through crowd-sourcing on top of the "MPII Cooking Composite Activities" video corpus (Rohrbach et al., 2012). In this paper, we selected two tasks "cutting cucumber" and "cutting bread" as our experimental data. Each task has 5 videos showing how different people perform the same task. Each video is segmented to a sequence of video clips where each video clip comes with one or more language descriptions. The original TACoS dataset does not contain annotation for grounded semantic roles.

Table 1: Statistics for a set of verbs and their semantic roles in our annotated dataset. The entry indicates the number of explicit/implicit roles for each category. "–" denotes no such role is observed in the data.[1]

Verb	Patient	Source	Destn	Location	Tool
take	251 / 0	102 / 149	2 / 248	–	–
put	94 / 0	–	75 / 19	–	–
get	247 / 0	62 / 190	0 / 239	–	–
cut	134 / 1	64 / 64	–	3 / 131	5 / 130
open	23 / 0	–	–	0 / 23	2 / 21
wash	93 / 0	–	–	26 / 58	2 / 82
slice	69 / 1	–	–	2 / 68	2 / 66
rinse	76 / 0	0 / 74	–	8 / 64	–
place	104 / 1	–	105 / 7	–	–
peel	29 / 0	–	–	1 / 27	2 / 27
remove	40 / 0	34 / 6	–	–	–

To support our investigation and evaluation, we had made a significant effort adding the following annotations. For each video clip, we annotated the objects' bounding boxes, their tracks, and their labels (cucumber, cutting_board, etc.) using VATIC (Vondrick et al., 2013). On average, each video clip is annotated with 15 tracks of objects. For each sentence, we annotated the ground truth parsing structure and the semantic frame for each verb. The ground truth parsing structure is the representation of dependency parsing results. The semantic frame of a verb includes slots, fillers, and their groundings. For each semantic role (including both explicit roles and implicit roles) of a given verb, we also annotated the ground truth grounding in terms of the object tracks and locations. In total, our annotated dataset includes 976 pairs of video clips and corresponding sentences, 1094 verbs occurrences, and 3593 groundings of semantic roles. To check annotation agreement, 10% of the data was annotated by two annotators. The kappa statistics is 0.83 (Cohen and others, 1960).

From this dataset, we selected 11 most frequent verbs (i.e., *get, take, wash, cut, rinse, slice, place, peel, put, remove, open*) in our current investigation for the following reasons. First, they are used more frequently so that we can have sufficient samples of each verb to learn the model. Second, they cover different types of actions: some are more related to the change of the state such as *take*, and some are more related to the process such as *wash*. As it turns out, these verbs also have different semantic role patterns as shown in Table 1. The `patient` roles of all these verbs are explicitly specified. This is not surprising as all these verbs are transitive verbs. There is a large variation for other roles. For example, for the verb *take*, the `destination` is rarely specified by lin-

[1]For some verbs (e.g., *get*), there is a slight discrepancy between the sum of implicit/explicit roles across different categories. This is partly due to the fact that some verb occurrences take more than one objects as grounding to a role. It is also possibly due to missed/duplicated annotation for some categories.

guistic expressions (i.e., only 2 instances), however it can be inferred from the video. For the verb *cut*, the `location` and the `tool` are also rarely specified by linguistic expressions. Nevertheless, these implicit roles contribute to the overall understanding of actions and should also be grounded too.

4.2 Automated Processing

To build the structure of the CRF as shown in Figure 2 and extract features for learning and inference, we have applied the following approaches to process language and vision.

Language Processing. Language processing consists of three steps to build a structure containing syntactic and semantic information. First, the Stanford Parser (Manning et al., 2014) is applied to create a dependency parsing tree for each sentence. Second, Senna (Collobert et al., 2011) is applied to identify semantic role labels for the key verb in the sentence. The linguistic entities with semantic roles are matched against the dependency nodes in the tree and the corresponding semantic role labels are added to the tree. Third, for each verb, the Propbank (Palmer et al., 2005) entries are searched to extract all relevant semantic roles. The implicit roles (i.e., not specified linguistically) are added as direct children of verb nodes in the tree. Through these three steps, the resulting tree from language processing has both explicit and implicit semantic roles. These trees are further transformed to the CRF structures based on a set of rules.

Vision Processing. A set of visual detectors are first trained for each type of objects. Here a random forest classifier is adopted. More specifically, we use 100 trees with HoG features (Dalal and Triggs, 2005) and color descriptors (Van De Weijer and Schmid, 2006). Both HoG and Color descriptors are used, because some objects are more structural, such as knives, human; some are more textured such as towels. With the learned object detectors, given a candidate video clip, we run the detectors at each 10th frame (less than 0.5 second), and find the candidate windows for which the detector score corresponding to the object is larger than a threshold (set as 0.5). Then using the detected window as a starting point, we adopt tracking-by-detection (Danelljan et al., 2014) to go forward and backward to track this object and obtain the candidate track with this object label.

Feature Extraction. Features in the CRF model can be divided into the following three categories:

1. *Linguistic features* include word occurrence and semantic role information. They are extracted by language processing.

2. *Track label features* are the label information for tracks in the video. The labels come from human annotation or automated visual processing depending on different experimental settings (described in Section 4.3).

3. *Visual features* are a set of features involving geometric relations between tracks in the video. One important feature is the histogram comparison score. It measures the similarity between distance histograms. Specifically, histograms of distance values between the tracks of the `predicate` and other roles for each verb are first extracted from the training video clips. For an incoming distance histogram, we calculate its Chi-Square distances (Zhang et al., 2007) from the pre-extracted training histograms with the same verb and the same role. its histogram comparison score is set to be the average of top 5 smallest Chi-Square distances. Other visual features include geometric information for single tracks and geometric relations between two tracks. For example, size, average speed, and moving direction are extracted for a single track. Average distance, size-ratio, and relative direction are extracted between two tracks. For features that are continuous, we discretized them into uniform bins.

To ground language into tracks from the video, instead of using track label features or visual features alone, we use a Cartesian product of these features with linguistic features. To learn the behavior of different semantic roles of different verbs, visual features are combined with the presence of both verbs and semantic roles through Cartesian product. To learn the correspondence between track labels and words, track label features are combined with the presence of words also through Cartesian product.

To train the model, we randomly selected 75% of annotated 976 pairs of video clips and corresponding sentences as training set. The remaining 25% were used as the testing set.

4.3 Experimental Setup

Comparison. To evaluate the performance of our approach, we compare it with two approaches.

- **Baseline**: To identify the grounding for each semantic role, the first baseline chooses the most possible track based on the object type conditional distribution given the verb and semantic role. If an object type corresponds to multiple tracks in the video, e.g., multiple drawers or knives, we then randomly select one of the tracks as grounding. We ran this baseline method five times and reported the average performance.

- **Tellex (2011)**: The second approach we compared with is based on an implementation (Tellex et al., 2011). The difference is that they don't explicitly model fine-grained semantic role information. For a better comparison, we map the grounding results from this approach to different explicit semantic roles according to the SRL annotation of the sentence. Note that this approach is not able to ground implicit roles.

More specifically, we compare these two approaches with two variations of our system:

- **GSRL$_{\text{wo_V}}$**: The CRF model using linguistic features and track label features (described in Section 4.2).

- **GSRL**: The full CRF model using linguistic features, track label features, and visual features(described in Section 4.2).

Configurations. Both automated language processing and vision processing are error-prone. To further understand the limitations of grounded SRL, we compare performance under different configurations along the two dimensions: (1) the CRF structure is built upon annotated ground-truth language parsing versus automated language parsing; (2) object tracking and labeling is based on annotation versus automated processing. These lead to four different experimental configurations.

Evaluation Metrics. For experiments that are based on annotated object tracks, we can simply use the traditional *accuracy* that directly measures the percentage of grounded tracks that are correct. However, for experiments using automated tracking, evaluation can be difficult as tracking itself poses significant challenges. The grounding results (to tracks) cannot be directly compared with the annotated ground-truth tracks. To address this problem, we have defined a new metric called *approximate accuracy*. This metric is motivated by previous computer vision work that evaluates tracking performance (Bashir and Porikli, 2006). Suppose the ground truth grounding for a role is track gt and the predicted grounding is track pt. The two tracks gt and pt are often not the same (although may have some overlaps). Suppose the number of frames in the video clip is k. For each frame, we calculate the distance between the centroids of these two tracks. If their distance is below a predefined threshold, we consider the two tracks overlap in this frame. We consider the grounding is correct if the ratio of the overlapping frames between gt and pt exceeds 50%. As can be seen, this is a lenient and an approximate measure of accuracy.

4.4 Results

The results based on the ground-truth language parsing are shown in Table 2, and the results based on automated language parsing are shown in Table 3. For results based on annotated object tracking, the performance is reported in *accuracy* and for results based on automated object tracking, the performance is reported in *approximate accuracy*. When the number of testing samples is less than 15, we do not show the result as it tends to be unreliable (shown as *NA*). Tellex (2011) does not address implicit roles (shown as "–"). The best performance score is shown in bold. We also conducted a two-tailed bootstrap significance testing (Efron and Tibshirani, 1994). The score with a "*****" indicates it is statistically significant ($p < 0.05$) compared to the baseline approach. The score with a "**+**" indicates

Table 2: Evaluation results based on annotated language parsing.

Accuracy On the Gold Recognition/Tracking Setting														
Methods	Predicate	Patient		Source		Destination		Location		Tool		Explicit All	Implicit All	All
		explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit			
Baseline	0.856	0.372	NA	0.225	0.314	0.311	0.569	NA	0.910	NA	0.853	0.556	0.620	0.583
Tellex(2011)	0.865	0.745	–	0.306	–	0.763	–	NA	–	NA	–	0.722	–	–
GSRL$_{\text{wo_V}}$	0.854	0.794$^{*}_{+}$	NA	0.375*	0.392$^{*}_{+}$	0.658*	0.615$^{*}_{+}$	NA	0.920$_{+}$	NA	0.793$_{+}$	0.768$^{*}_{+}$	0.648$^{*}_{+}$	0.717*
GSRL	**0.878**$^{*}_{+}$	**0.839**$^{*}_{+}$	NA	**0.556**$^{*}_{+}$	**0.684**$^{*}_{+}$	**0.789***	**0.641**$^{*}_{+}$	NA	**0.930**$_{+}$	NA	**0.897**$^{*}_{+}$	**0.825**$^{*}_{+}$	**0.768**$^{*}_{+}$	**0.8***
Approximated Accuracy On the Automated Recognition/Tracking Setting														
Methods	Predicate	Patient		Source		Destination		Location		Tool		Explicit All	Implicit All	All
		explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit			
Baseline	0.529	0.206	NA	0.169	0.119	0.236	0.566	NA	**0.476**	NA	0.6	0.352	0.393	0.369
Tellex(2011)	**0.607**	0.233	–	0.154	–	0.333	–	NA	–	NA	–	0.359	–	–
GSRL$_{\text{wo_V}}$	0.582*	0.244*	NA	**0.262**$^{*}_{+}$	**0.126**$_{+}$	**0.485**$^{*}_{+}$	**0.613**$^{*}_{+}$	NA	0.467$_{+}$	NA	**0.714**$^{*}_{+}$	**0.410**$^{*}_{+}$	**0.425**$^{*}_{+}$	**0.417***
GSRL	0.548	**0.263***	NA	**0.262**$^{*}_{+}$	0.086$_{+}$	0.394*	0.514$_{+}$	NA	0.456$_{+}$	NA	0.688$^{*}_{+}$	0.399$^{*}_{+}$	0.381$_{+}$	0.391*
Upper_Bound	0.920	0.309	NA	0.277	0.252	0.636	0.829	NA	0.511	NA	0.818	0.577	0.573	0.575

Table 3: Evaluation results based on automated language parsing.

Accuracy On the Gold Recognition/Tracking Setting														
Methods	Predicate	Patient		Source		Destination		Location		Tool		Explicit All	Implicit All	All
		explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit			
Baseline	0.881	0.318	NA	0.203	0.316	0.235	0.607	NA	0.877	NA	**0.895**	0.539	0.595	0.563
Tellex(2011)	**0.903**	0.746	–	0.156	–	0.353	–	NA	–	NA	–	0.680	–	–
GSRL$_{\text{wo_V}}$	0.873	0.813$^{*}_{+}$	NA	0.328$^{*}_{+}$	0.360$_{+}$	**0.412***	0.648$^{*}_{+}$	NA	0.877$_{+}$	NA	0.818$_{+}$	0.769$^{*}_{+}$	0.611$_{+}$	0.7*
GSRL	0.873	**0.875**$^{*}_{+}$	NA	**0.453**$^{*}_{+}$	**0.667**$^{*}_{+}$	**0.412***	**0.667**$^{*}_{+}$	NA	**0.891**$_{+}$	NA	0.891$_{+}$	**0.823**$^{*}_{+}$	**0.741**$^{*}_{+}$	**0.787***
Approximated Accuracy On the Automated Recognition/Tracking Setting														
Methods	Predicate	Patient		Source		Destination		Location		Tool		Explicit All	Implicit All	All
		explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit	explicit	implicit			
Baseline	0.543	0.174	NA	0.121	0.113	0.093	0.594	NA	**0.612**	NA	0.567	0.327	0.405	0.362
Tellex(2011)	0.598	0.218	–	0.086	–	0.00	–	NA	–	NA	–	0.322	–	–
GSRL$_{\text{wo_V}}$	**0.618***	**0.243***	NA	**0.190**$^{*}_{+}$	**0.120**$_{+}$	**0.133**$_{+}$	**0.641**$^{*}_{+}$	NA	0.585$_{+}$	NA	**0.723**$^{*}_{+}$	**0.401**$^{*}_{+}$	**0.434**$^{*}_{+}$	**0.415***
GSRL	0.493	**0.243***	NA	**0.190**$^{*}_{+}$	0.063$_{+}$	**0.133**$_{+}$	0.612$_{+}$	NA	0.554$_{+}$	NA	0.617$_{+}$	0.367$^{*}_{+}$	0.386$_{+}$	0.375
Upper_Bound	0.908	0.277	NA	0.259	0.254	0.4	0.854	NA	0.631	NA	0.830	0.543	0.585	0.561

it is statistically significant ($p < 0.05$) compared to the approach (Tellex et al., 2011).

For experiments based on automated object tracking, we also calculated an *upper_bound* to assess the best possible performance which can be achieved by a perfect grounding algorithm given the current vision processing results. This *upper_bound* is calculated based on grounding each role to the track which is closest to the ground-truth annotated track. For the experiments based on annotated tracking, the *upper_bound* would be 100%. This measure provides some understandings about how good the grounding approach is given the limitation of vision processing. Notice that the grounding results in the gold/automatic language processing setting are not directly comparable as the automatic SRL can misidentify frame elements.

4.5 Discussion

As shown in Table 2 and Table 3, our approach consistently outperforms the baseline (for both explicit and implicit roles) and the Tellex (2011) approach. Under the configuration of gold recognition/tracking, the incorporation of visual features further improves the performance. However, this performance gain is not observed when automated object tracking and labeling is used. One possible explanation is that as we only had limited data, we did not use separate data to train models for

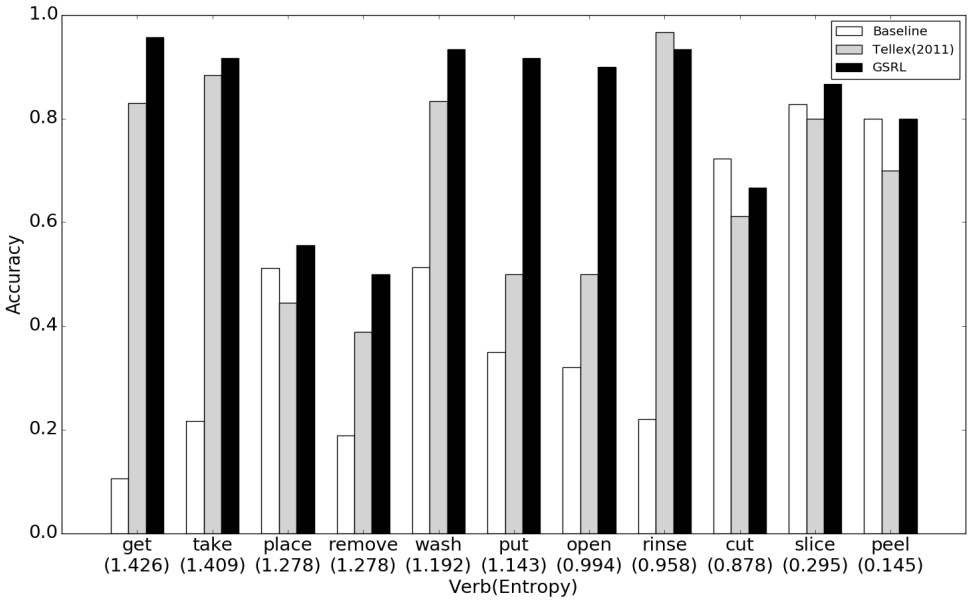

Figure 3: The relation between the accuracy and the entropy of each verb's patient from the gold language, gold visual recognition/tracking setting. The entropy for the patient role of each verb is shown below the verb.

object recognition/tracking. So the GSRL model was trained with gold recognition/tracking data and tested with automated recognition/tracking data.

By comparing our method with Tellex (2011), we can see that by incorporating fine grained semantic role information, our approach achieves better performance on almost all the explicit role (except for the `patient` role under the automated tracking condition).

The results have also shown that some roles are easier to ground than others in this domain. For example, the `predicate` role is grounded to the hand tracks (either left hand or right hand), there are not many variations such that the simple baseline can achieve pretty high performance, especially when annotated tracking is used. The same situation happens to the `location` role as most of the locations happen near the *sink* when the verb is *wash*, and near the *cutting board* for verbs like *cut*, etc. However, for the `patient` role, there is a large difference between our approach and baseline approaches as there is a larger variation of different types of objects that can participate in the role for a given verb.

For experiments with automated tracking, the *upper_bound* for each role also varies. Some roles (e.g., `patient`) have a pretty low upper bound. The accuracy from our full GSRL model is already quite close to the upper bound. For other roles such as `predicate` and `destination`, there is a larger gap between the current performance and the upper bound. This difference reflects the model's capability in grounding different roles.

Figure 3 shows a close-up look at the grounding performance to the `patient` role for each verb under the gold parsing and gold tracking configuration. The reason we only show the results of `patient` role here is every verb has this role to be grounded. For each verb, we also calculated its entropy based on the distribution of different types of objects that can serve as the `patient` role in the training data. The entropy is shown at the bottom of the figure. For verbs such as *take* and *put*, our full GSRL model leads to much better performance compared to the baseline. As the baseline approach relies on the entropy of the potential grounding for a role, we further measured the improvement of the performance and calculated the correlation between the improvement and the entropy of each verb. The result shows that Pearson coefficient between the entropy and the improvement of GSRL over the baseline is 0.614. This indicates the improvement from GSRL is positively correlated with the entropy value associated with a role, implying the GSRL model can deal with

more uncertain situations. For the verb *cut*, The GSRL model performs slightly worse than the baseline. One explanation is that the possible objects that can participate as a patient for *cut* are relatively constrained where simple features might be sufficient. A large number of features may introduce noise, and thus jeopardizing the performance.

We further compare the performance of our full GRSL model with Tellex (2011) (also shown in Figure 3) on the `patient` role of different verbs. Our approach outperforms Tellex (2011) on most of the verbs, especially *put* and *open*. A close look at the results have shown that in those cases, the `patient` roles are often specified by pronouns. Therefore, the track label features and linguistic features are not very helpful, and the correct grounding mainly depends on visual features. Our full GSRL model can better capture the geometry relations between different semantic roles by incorporating fine-grained role information.

5 Conclusion and Future Work

This paper investigates a new problem on grounded semantic role labeling. Besides semantic roles explicitly mentioned in language descriptions, our approach also grounds implicit roles which are not explicitly specified. As implicit roles also capture important participants related to an action (e.g., tools used in the action), our approach provides a more complete representation of action semantics which can be used by artificial agents for further reasoning and planning towards the physical world. Our empirical results on a complex cooking domain have shown that, by incorporating semantic role information with visual features, our approach can achieve better performance compared to baseline approaches. Our results have also shown that grounded semantic role labeling is a challenging problem which often depends on the quality of automated visual processing (e.g., object tracking and recognition).

There are several directions for future improvement. First, the current alignment between a video clip and a sentence is generated by some heuristics which are error-prone. One way to address this is to treat alignment and grounding as a joint problem. Second, our current visual features have not shown effective especially when they are extracted based on automatic visual processing. This is partly due to the complexity of the scene from the TACoS dataset and the lack of depth information. Recent advances in object tracking algorithms (Yang et al., 2013; Milan et al., 2014) together with 3D sensing can be explored in the future to improve visual processing. Moreover, linguistic studies have shown that action verbs such as *cut* and *slice* often denote some change of state as a result of the action (Hovav and Levin, 2010; Hovav and Levin, 2008). The change of state can be perceived from the physical world. Thus another direction is to systematically study causality of verbs. Causality models for verbs can potentially provide top-down information to guide intermediate representations for visual processing and improve grounded language understanding.

The capability of grounding semantic roles to the physical world has many important implications. It will support the development of intelligent agents which can reason and act upon the shared physical world. For example, unlike traditional action recognition in computer vision (Wang et al., 2011), grounded SRL will provide deeper understanding of the activities which involve participants in the actions guided by linguistic knowledge. For agents that can act upon the physical world such as robots, grounded SRL will allow the agents to acquire the grounded structure of human commands and thus perform the requested actions through planning (e.g., to follow the command "put the cup on the table"). Grounded SRL will also contribute to robot action learning where humans can teach the robot new actions (e.g., simple cooking tasks) through both task demonstration and language instruction.

6 Acknowledgement

The authors are grateful to Austin Littley and Zach Richardson for their help on data annotation, and to anonymous reviewers for their valuable comments and suggestions. This work was supported in part by IIS-1208390 from the National Science Foundation and by N66001-15-C-4035 from the DARPA SIMPLEX program.

References

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *TACL*, 1:49–62.

Faisal Bashir and Fatih Porikli. 2006. Performance evaluation of object detection and tracking systems. In *Proceedings 9th IEEE International Workshop on PETS*, pages 7–14.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47.

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*.

David L Chen and Raymond J Mooney. 2008. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135. ACM.

Jacob Cohen et al. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost van de Weijer. 2014. Adaptive color attributes for real-time visual tracking. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1090–1097. IEEE.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Desmond Elliott and Arjen de Vries. 2015. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 42–52, Beijing, China, July. Association for Computational Linguistics.

Bastianelli Emanuele, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. 2013. Textual inference and meaning representation in human robot interaction. In *Joint Symposium on Semantic Processing.*, page 65.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Malka Rappaport Hovav and Beth Levin. 2008. Reflections on manner/result complementarity. *Lecture notes*.

Malka Rappaport Hovav and Beth Levin. 2010. Reflections on Manner / Result Complementarity. *Lexical Semantics, Syntax, and Event Structure*, pages 21–38.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. June.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October. Association for Computational Linguistics.

Jayant Krishnamurthy and Thomas Kollar. 2013. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206.

Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. 2013. Generalizing image captions for image-text parallel corpus. In *ACL (2)*, pages 790–796. Citeseer.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Changsong Liu and Joyce Y. Chai. 2015. Learning to mediate perceptual differences in situated human-robot dialogue. In *The Twenty-Ninth Conference on Artificial Intelligence (AAAI-15)*. to appear.

Changsong Liu, Rui Fang, and Joyce Chai. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149, Seoul, South Korea.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Anton Milan, Stefan Roth, and Kaspar Schindler. 2014. Continuous energy minimization for multitarget tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):58–72.

Iftekhar Naim, Young C. Song, Qiguang Liu, Liang Huang, Henry Kautz, Jiebo Luo, and Daniel Gildea. 2015. Discriminative unsupervised alignment of natural language instructions with corresponding video segments. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 164–174, Denver, Colorado, May–June. Association for Computational Linguistics.

Luis Gilberto Mateos Ortiz, Clemens Wolff, and Mirella Lapata. Learning to interpret and describe abstract scenes. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1505–1515.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Vignesh Ramanathan, Percy Liang, and Li Fei-Fei. 2013. Video event understanding using natural language descriptions. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 905–912. IEEE.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics (TACL)*, 1:25–36.

Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele. 2012. Script data for attribute-based recognition of composite activities. In *Computer Vision–ECCV 2012*, pages 144–157. Springer.

Deb Roy. 2005. Grounding words in perception and action: computational insights. *TRENDS in Cognitive Sciences*, 9(8):389–396.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *EMNLP-CoNLL*, pages 12–21.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*.

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.

Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. 2014. Joint video and text parsing for understanding events and answering queries. *MultiMedia, IEEE*, 21(2):42–70.

Joost Van De Weijer and Cordelia Schmid. 2006. Coloring local feature extraction. In *Computer Vision–ECCV 2006*, pages 334–348. Springer.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, Denver, Colorado, May–June. Association for Computational Linguistics.

Carl Vondrick, Donald Patterson, and Deva Ramanan. 2013. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204.

Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE.

Yezhou Yang, Cornelia Fermuller, and Yiannis Aloimonos. 2013. Detection of manipulation action consequences (mac). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2563–2570.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *ACL (1)*, pages 53–63.

Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73(2):213–238.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China, July. Association for Computational Linguistics.