Commonly Uncommon

Semantic Sparsity in Situation Recognition

Saeid Naderiparizi
Department of Computer Science

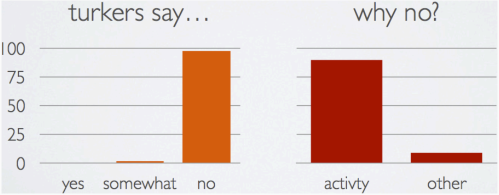
Xiaomeng Ju
Department of Statistics

Table of contents

Background

What is Situation Recognition?

Is the same thing happening in these images?

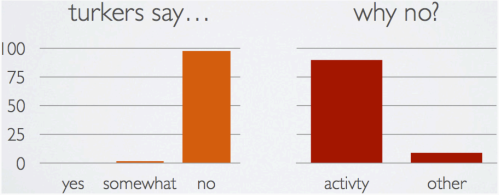


slide from Mark Yatskar

What is Situation Recognition?

Is the same thing happening in these images?

Activity



slide from Mark Yatskar

What is Situation Recognition?

Is the same thing happening in these images?

Activity

slide from Mark Yatskar

What is Situation Recognition?

Is the same thing happening in these images?

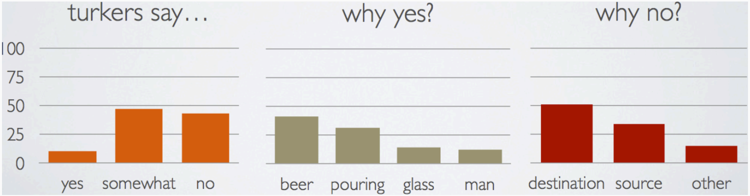
Activity
Objects

slide from Mark Yatskar

What is Situation Recognition?

Is the same thing happening in these images?

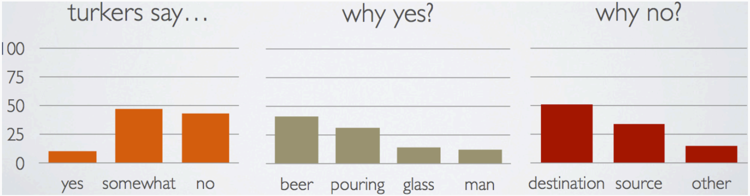
Activity Objects



slide from Mark Yatskar

Is the same thing happening in these images?

Activity
Objects
Roles



slide from Mark Yatskar

What is Situation Recognition?

What is happening in an image?

A man is carrying a baby on his chest outdoors.

carrying			
agent	item	agentpart	place
man	baby	chest	outdoors

slide from Mark Yatskar

How to extract roles?

FrameNet: a semantic-role-labeling project.

- The meanings of most words can best be understood on the basis of a **semantic frame**.

cooking					
agent	food	container	heatsource	tool	place
noun	noun	noun	noun	noun	noun

- For a frame f,
    - Set of semantic roles is called E_f.
    - Set of pairs of semantic roles and their values is called a ”realized frame” R_f.

Problem Formulation

- A situation $S = (v, R_f)$, where v is a verb and R_f is a realized frame.
- Each element in R_f is (e, n_e), is a pair of semantic role e and a noun n_e.

$$(carrying, \{(agent, man), (item, table), (agentpart, back), (place, street)\})$$

- Frame f

$$\{(agent,), (item,), (agentpart,), (place,)\}$$

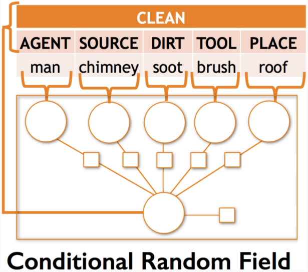
- A verb $v \in V$ is mapped to exactly one frame $f \in F$ that is described with a set of semantic roles.
- V and F are derived from FrameNet (Fillmore et al. 2003)
- Situation recognition:

$$\underset{S}{\operatorname{argmax}} P(S|i)$$

Conditional Random Field (CRF): Basics

- CRF is a probabilistic graphical model that fits the conditional distributions $P(Y|X)$. In our setting $P(S|i)$.
- Conditional distribution is factorized using **potentials** defined on subsets of Y:

$$P(Y|X) \propto \psi_1(D_1, X)\psi_2(D_2, X)...$$



For situation recognition:

$$P(S|i;\theta) \propto \psi_v(v, i; \theta) \prod_{(e,n_e)\in R_f} \psi_e(v, e, n_e, i; \theta)$$

Conditional Random Field (CRF): Potentials

- $$P(S|i;\theta) \propto \psi_v(v,i;\theta) \prod_{(e,n_e)\in R_f} \psi_e(v,e,n_e,i;\theta)$$
- Verb potential:
$$\psi_v(v,i;\theta) = e^{\phi_v(v,i;\theta)}$$
- Verb-Role-None potential:
$$\psi_e(v,e,n_e,i;\theta) = e^{\phi_e(v,e,n_e,i;\theta)}$$

Conditional Random Field (CRF): Architecture of Previous Work

- Let $g_i \in \mathbb{R}^p$ be an image representation from VGG

$$\phi_v(v, i; \theta) = g_i^T \theta_v, \qquad \phi_e(v, e, n_e, i; \theta) = g_i^T \theta_{v,e,n_e}.$$

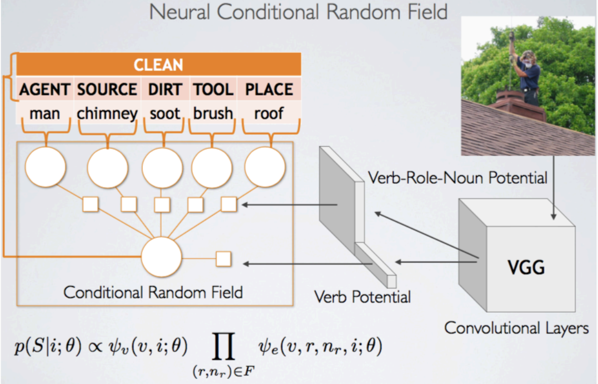


Figure 7: Situation recognition: visual semantic role labeling for image understanding (Yatskar et al. 2016)

Conditional Random Field (CRF): Training

- Training data: $\{\text{image}_i, S \in A_i\}_{i=1}^{n}$ (A_i ground truth situations)
- Optimize the log-likelihood of observing at least one situation $S \in A_i$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log \left(1 - \prod_{S \in A_i} (1 - p(S|i;\theta) \right)$$

Potential problem:

- $\phi_e(v, e, n_e, i; \theta) = g_i^T \theta_{v,e,n_e}$, need to compute θ_{v,e,n_e} for every combination of (v, e, n_e).
- Hard to obtain an accurate estimate with rare (v, e, n_e) combinations.

Motivation

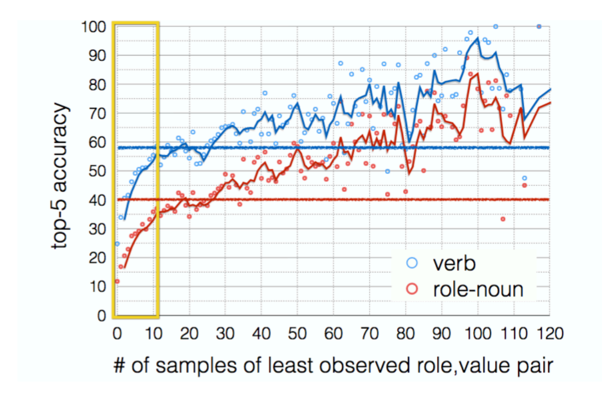
Motivation: Semantic sparsity

- Semantic sparsity: “there are a combinatorial number of possible outputs, no dataset can cover them all” (Yatskar, 2016)
- For a given verb, many role-value combinations are rare.

CARRYING

ROLE	VALUE	ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	WOMAN	AGENT	MAN
ITEM	BABY	ITEM	BUCKET	ITEM	TABLE
AGENTPART	CHEST	AGENTPART	HEAD	AGENTPART	BACK
PLACE	OUTSIDE	PLACE	PATH	PLACE	STREET

- Semantic sparsity is common: 35% of the (verb, role, noun) pairs appeared less than 10 times in the training set.
- Current CRF model performs badly with rarely observed role-value pairs.



Contribution of the paper

The paper “Commonly Uncommon” (Yatskar et al, 2016) deals with semantic sparsity in situation recognition by

- (1) introducing compositional CRF that shares information of the nouns between roles.
- (2) semantically augmenting the training data with gathered web data.

Methodology

Compositional CRF: Basic Idea

JUMPING

ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

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

- Some nouns are shared across different roles (e.g. water)
- Independent representation of noun, (verb,role), and image.

Compositional CRF: Tensor Potential

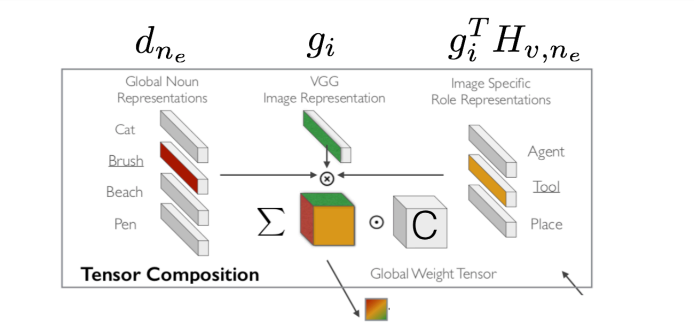
- CRF:

$$\phi_e(v, e, n_e, i; \theta) = g_i^T \theta_{v,e,n_e}.$$

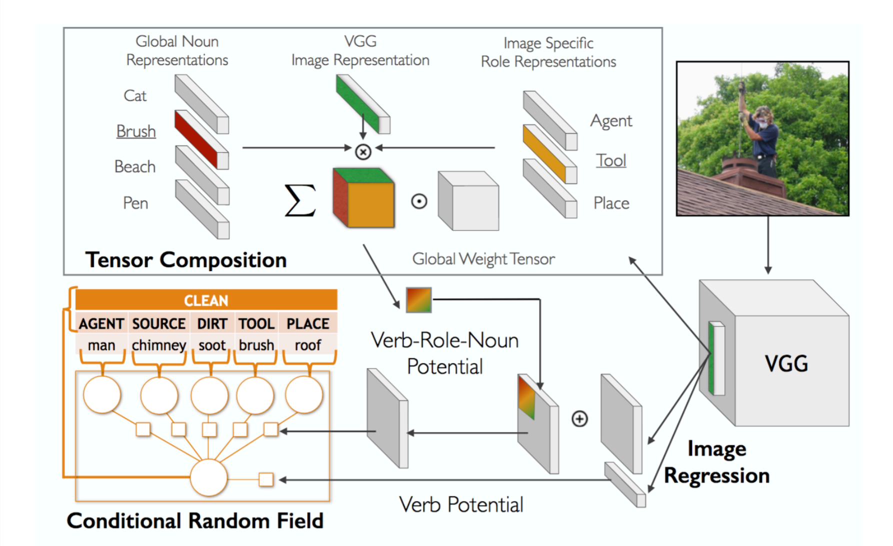
- Compositional CRF

$$T(v, e, n_e, g_i) = C \odot (d_{n_e} \otimes g_i^T H_{v,e} \otimes g_i)$$

$$\phi_e(v, e, n_e, i; \theta) = \sum_{x=0}^{M} \sum_{y=0}^{O} \sum_{z=0}^{P} T(v, e, n_e, g_i)[x, y, z].$$



Compositional CRF: Proposed Architecture



Semantic data augmentation: Overview

- Generate descriptive sentences.
- Use image search to find images for data augmentation.
- Pre-train the network on images from the web.
- Use ”Self Training” to reduce effect of noise.

Semantic data augmentation: Terminology

We only do data augmentation for ”uncommon situations”:

For each image i, the groundtruth situation is $S_i = (v_i, R_{f_i})$.

S is an uncommons situation if $\exists (e, n_e) \in R_f : \#\{(e, n_e) \in R_{f_i} | v_i = v\}$ is small.

Semantic data augmentation: Generate descriptive sentences

For an uncommon situation $S = (v, R_f)$, enumerate all sub-pieces of R_f.

Example:

$$R_f = (carrying, \{(agent, man), (item, table), (agentpart, back), (place, street)\})$$

$$\Big\downarrow$$

$$(carrying, \{(agent, man)\}$$

$$(carrying, \{(agent, man), (item, table)\}$$

$$(carrying, \{(item, table)\}$$

$$\vdots$$

Semantic data augmentation: Generate descriptive sentences

Using a template for each verb, each sub-structure is deterministically converted into a phrase.

Example:
`{agent}` carrying `{item}` {with `agentpart`} {in `place`}

$$\downarrow$$

man carrying
man carrying table

Semantic data augmentation: Retrieve images

- Generated phrases are used as queries to Google image search.
- Construct a set of images annotated with a verb and partially complete realized frames.

Semantic data augmentation: Pre-training

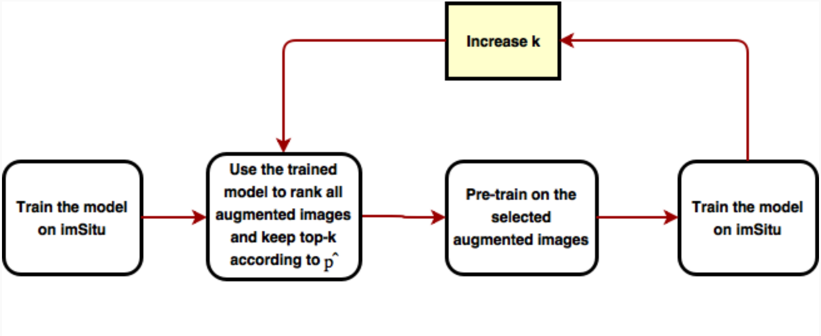
- Retrieved images are annotated partially.
- Partially realized frame: R_{pf}
- Use marginal likelihood for computing potentials.

$$\hat{p}(S|i;\theta) \propto \psi_v(v,i;\theta) \prod_{(e,n_e)\in R_{pf}} \psi_e(v,e,n_e,i;\theta) \times \prod_{e\notin R_{pf} \wedge e\in E_f} \sum_{n\in N} \psi_e(v,e,n,i;\theta)$$

carrying			
agent	item	agentpart	place
man	-	-	-

Semantic data augmentation: Self Training

Retrieved images are noisy.



Experimental setup and results

Experimental setup: Baselines

1. Image Regression (Yatskar et al, 2016)

$$\phi_e(v, e, n_e, i, \theta) = g_i^T \theta_{v,e,n_e}$$

2. Noun potential

$$p(S|\theta_i) = \psi_v(v, i, ; \theta) \prod_{(e,n_e) \in R_f} \psi_e(v, e, n_e, i; \theta) \psi_{n_e}(n_e, i; \theta)$$

3. Inner product composition

$$\phi_e(v, e, n_e, i) = \sum_k d_{n_e}^T H_{(k,v,e)} g_i$$

Results

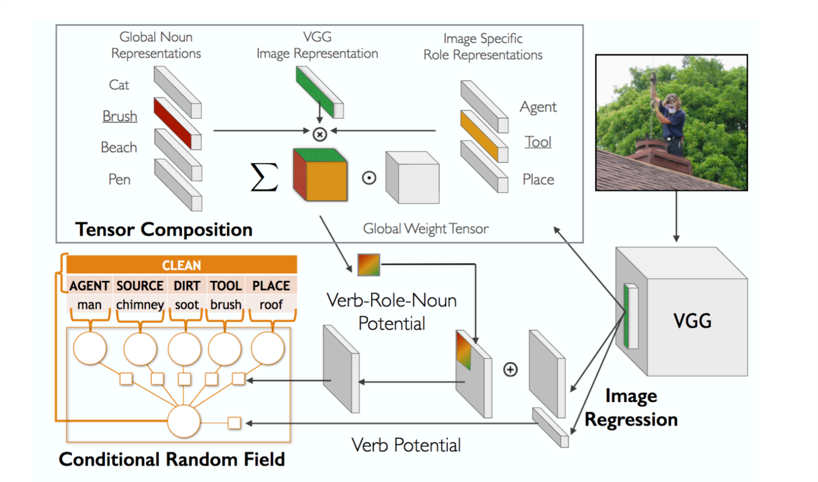
			top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
			verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	1	Baseline: Image Regression [44]	32.25	24.56	14.28	58.64	42.68	22.75	65.90	29.50	36.32
	2	Noun Potential + reg	27.64	21.21	12.21	53.95	39.95	21.45	68.87	32.31	34.70
	3	Inner product composition + reg	32.13	24.77	14.71	58.33	42.93	23.14	66.79	30.2	36.62
	4	Tensor composition	31.73	24.04	13.73	58.06	42.64	22.7	68.73	32.14	36.72
	5	Tensor composition + reg	32.91	25.39	14.87	59.92	44.5	24.04	69.39	33.17	38.02
+ SA	6	Baseline : Image Regression	32.40	24.14	15.17	59.10	44.04	24.40	68.03	31.93	37.53
	7	Tensor composition + reg	34.04	26.47	**15.73**	61.75	46.48	**25.77**	**70.89**	**35.08**	39.53
	8	Tensor composition + reg + self train	**34.20**	**26.56**	15.61	**62.21**	**46.72**	25.66	70.80	34.82	**39.57**

Results on the full imSitu development set

			top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
			verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	1	Baseline: image regression [44]	19.89	11.68	**2.85**	44.00	24.93	**6.16**	50.80	9.97	19.92
	2	Noun potential + reg	15.88	9.13	1.86	38.22	22.28	5.46	54.65	11.91	19.92
	3	Inner product composition + reg	18.96	10.69	1.89	42.53	23.28	3.69	49.54	6.46	19.63
	4	Tensor composition	19.78	11.28	2.26	42.66	24.42	5.57	54.06	11.47	21.43
	5	Tensor composition + reg	**21.12**	11.89	2.20	45.14	25.51	5.36	53.58	10.62	21.93
+ SA	6	Baseline : image regression	19.95	11.44	2.13	43.08	24.56	4.95	51.55	8.41	20.76
	7	Tensor composition + reg	20.08	11.58	2.22	44.82	26.02	5.55	55.45	11.53	22.16
	8	Tensor composition + reg + self train	20.52	**11.91**	2.34	**45.94**	**26.99**	6.06	**55.90**	**12.04**	**22.71**

Results on the rare portion of imSitu development set

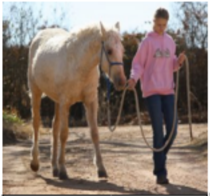
Results


Global Noun Representations
VGG Image Representation
Image Specific Role Representations
Cat
Brush
Beach
Pen
Agent
Tool
Place
Tensor Composition
Global Weight Tensor
CLEAN
AGENT
SOURCE
DIRT
TOOL
PLACE
man
chimney
soot
brush
roof
Verb-Role-Noun Potential
VGG
Image Regression
Verb Potential
Conditional Random Field

Results

DOUSING	
ROLE	VALUE
AGENT	PERSON
LIQUID	WATER
DEST.	FIRE
PLACE	BUILDING

FLOATING	
ROLE	VALUE
AGENT	PERSON
MEDIUM	WATER
TOOL	∅
PLACE	OUTSIDE



LEADING	
ROLE	VALUE
AGENT	WOMAN
FOLLOWER	HORSE
PLACE	ROAD

RIDING	
ROLE	VALUE
AGENT	TRUCK
VEHICLE	HORSE
PLACE	FIELD

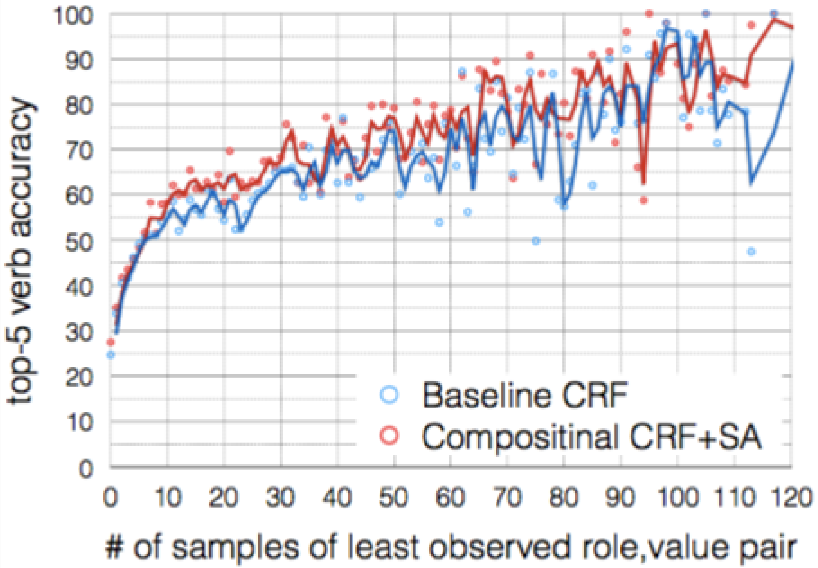
		top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
		verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	Baseline: Image Regression [44]	32.34	24.64	14.19	58.88	42.76	22.55	65.66	28.96	36.25
	Tensor composition + reg	32.96	25.32	14.57	60.12	44.64	24.00	69.2	32.97	37.97
+ SA	Baseline : Image Regression	32.3	24.95	14.77	59.52	44.08	23.99	67.82	31.46	37.36
	Tensor composition + reg + self train	**34.12**	**26.45**	**15.51**	**62.59**	**46.88**	**25.46**	**70.44**	**34.38**	**39.48**

Results on the rare portion of imSitu test set

		top-1 predicted verb			top-5 predicted verbs			ground truth verbs		
		verb	value	value-all	verb	value	value-all	value	value-all	mean
imSitu	Baseline: Image Regression [44]	**20.61**	11.79	**3.07**	44.75	24.85	5.98	50.37	9.31	21.34
	Tensor composition + reg	19.96	11.57	2.30	44.89	25.26	4.87	53.39	10.15	21.55
+ SA	Baseline : Image Regression	19.46	11.15	2.13	43.52	24.14	4.65	51.21	8.26	20.57
	Tensor composition + reg + self train	20.32	**11.87**	2.52	**47.07**	**27.50**	**6.35**	**55.72**	**12.28**	**22.95**

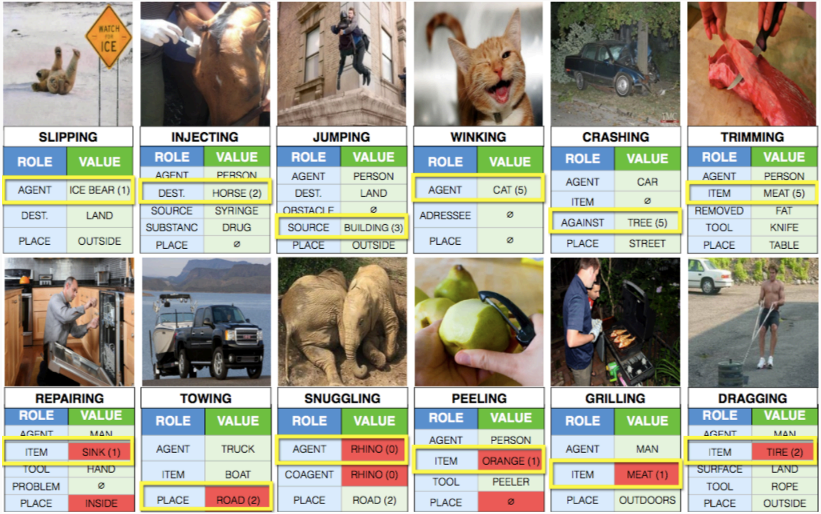
Results on the rare portion of imSitu test set

Results



top-5 verb accuracy
100
90
80
70
60
50
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
110
120
# of samples of least observed role,value pair
Baseline CRF
Compositinal CRF+SA

Results


WATCH FOR ICE
SLIPPING
ROLE VALUE
AGENT ICE BEAR (1)
DEST. LAND
PLACE OUTSIDE
INJECTING
ROLE VALUE
AGENT PERSON
DEST. HORSE (2)
SOURCE SYRINGE
SUBSTANC DRUG
PLACE ⌀
JUMPING
ROLE VALUE
AGENT PERSON
DEST. LAND
OBSTACLE ⌀
SOURCE BUILDING (3)
PLACE OUTSIDE
WINKING
ROLE VALUE
AGENT CAT (5)
ADRESSEE ⌀
PLACE ⌀
CRASHING
ROLE VALUE
AGENT CAR
ITEM ⌀
AGAINST TREE (5)
PLACE STREET
TRIMMING
ROLE VALUE
AGENT PERSON
ITEM MEAT (5)
REMOVED FAT
TOOL KNIFE
PLACE TABLE
REPAIRING
ROLE VALUE
AGENT MAN
ITEM SINK (1)
TOOL HAND
PROBLEM ⌀
PLACE INSIDE
TOWING
ROLE VALUE
AGENT TRUCK
ITEM BOAT
PLACE ROAD (2)
SNUGGLING
ROLE VALUE
AGENT RHINO (0)
COAGENT RHINO (0)
PLACE ROAD (2)
PEELING
ROLE VALUE
AGENT PERSON
ITEM ORANGE (1)
TOOL PEELER
PLACE ⌀
GRILLING
ROLE VALUE
AGENT MAN
ITEM MEAT (1)
PLACE OUTDOORS
DRAGGING
ROLE VALUE
AGENT MAN
ITEM TIRE (2)
SURFACE LAND
TOOL ROPE
PLACE OUTSIDE

Future Work

Future Work

- Follow-up publications
  - Li et al. (2017) captures joint dependencies between roles using neural networks defined on a graph.
  - Mallya and Lazebnik (2017) proposes Recurrent Neural Network (RNN) models to predict structured 'image situations'.
- Our thoughts:
  - Multiple frames corresponding to a given verb.
  - Predict the number of situations and their realizations for a given image.
  - A generalized definition of situation. (not only defined with (v, R_f)).

References

Yatskar, Mark, Vicente Ordonez, Luke Zettlemoyer, and Ali Farhadi.
Commonly uncommon: Semantic sparsity in situation recognition.
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017.

Yatskar, Mark, Luke Zettlemoyer, and Ali Farhadi.
Situation recognition: Visual semantic role labeling for image understanding.
Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

References

Mallya, Arun, and Svetlana Lazebnik
Recurrent models for situation recognition.
arXiv preprint arXiv:1703.06233 (2017)

Li, Ruiyu, Makarand Tapaswi, Renjie Liao, Jiaya Jia, Raquel Urtasun, and Sanja Fidler.
Situation recognition with graph neural networks.
arXiv preprint arXiv:1708.04320 (2017)

Thank you