Making TACoS: Grounding Distributional Models of Action Descriptions in Videos

Göttingen Symposium on "The Semantics of Action"

10. Juni 2013

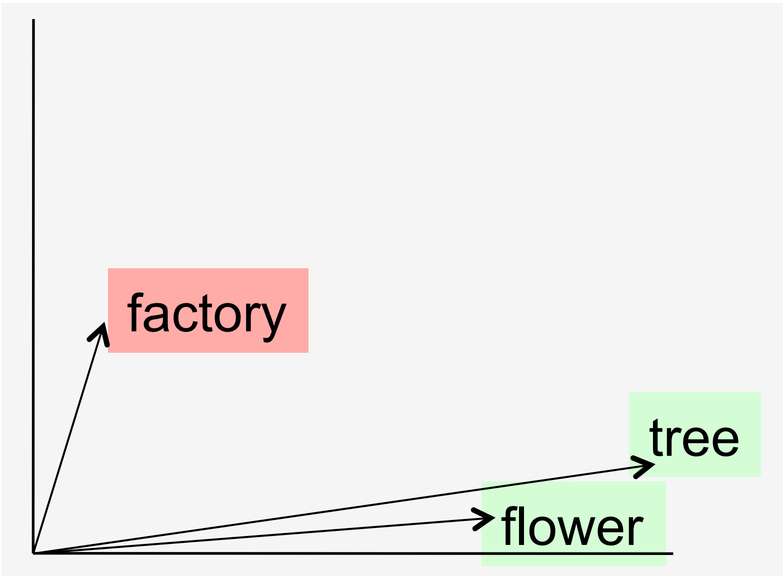
Manfred Pinkal, Saarland University

Distributional Hypothesis

Words that occur in the same contexts tend to have similar meanings.

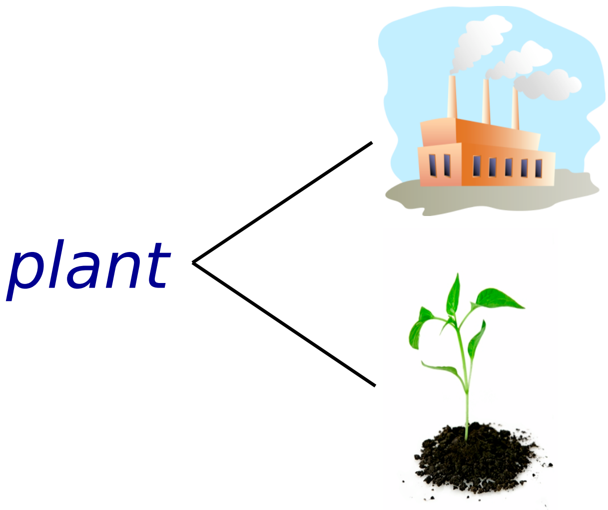
Distributional Similarity

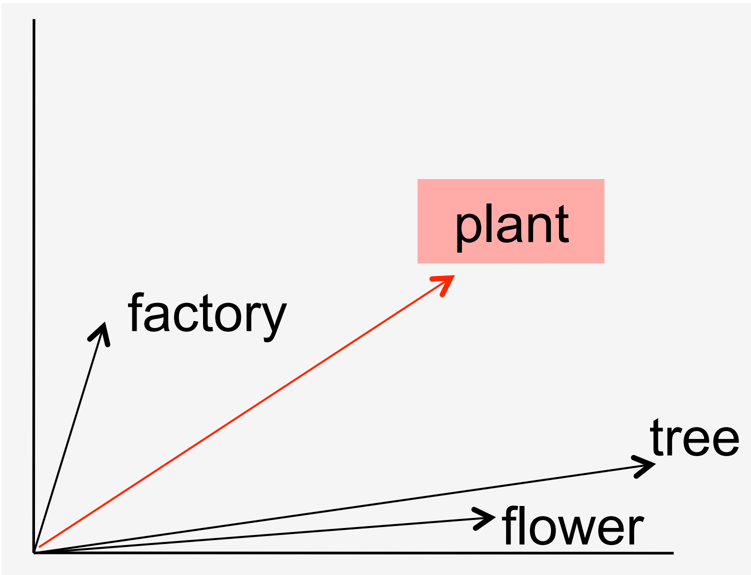
$$sim\ (a,b) = \cos\left(\vec{a},\vec{b}\right)$$



	factory	flower	tree	water	fork
...	...	...	...	...	...
grow	15	147	330	106	3
garden	5	200	198	118	17
worker	279	0	5	18	0
production	102	6	9	28	0
wild	3	216	35	30	0
...	...	...	...	...	...

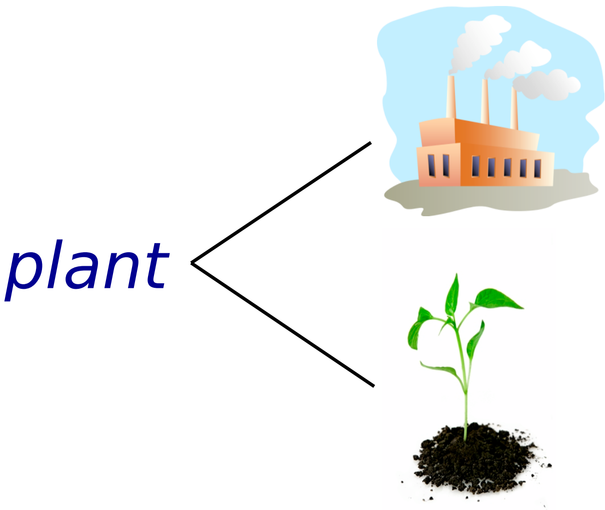
Contextual Specification



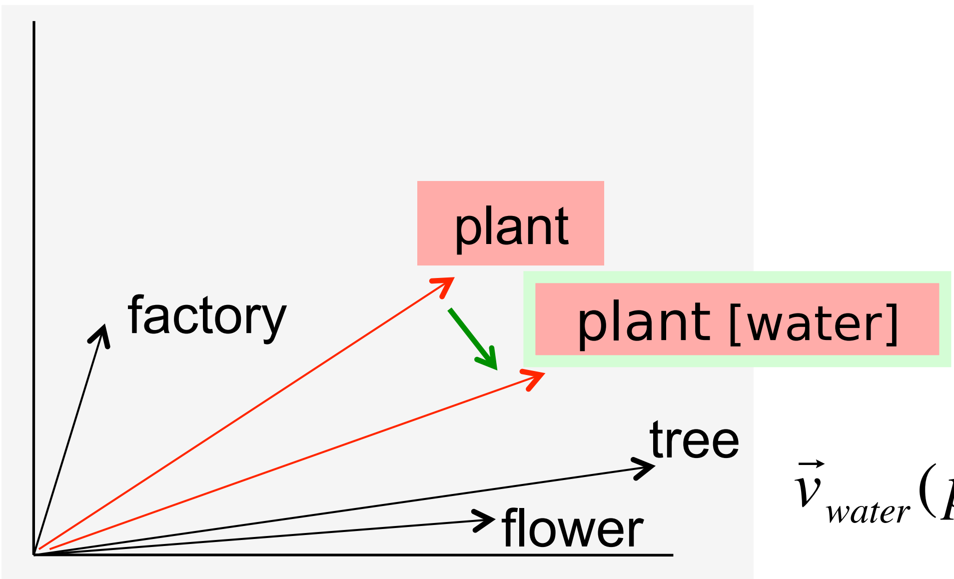


	plant	factory	flower	tree	water	fork
...	...	...	...	...	...	...
grow	517	15	147	330	106	3
garden	316	5	200	198	118	17
worker	84	279	0	5	18	0
production	130	102	6	9	28	0
wild	96	3	216	35	30	0
...	...	...	...	...	...	...

Contextual Specification



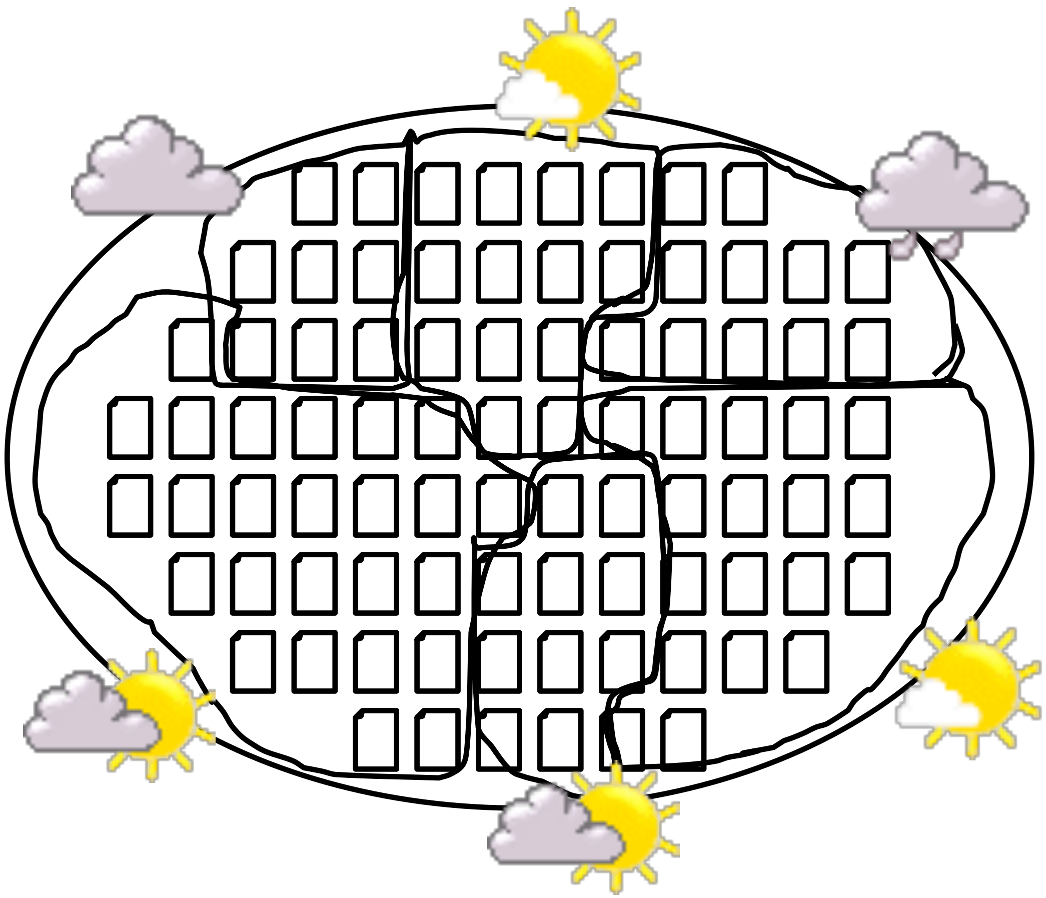
	plant	factory	flower	tree	water	fork
...	...	...	...	...	...	...
grow	517	15	147	330	106	3
garden	316	5	200	198	118	17
worker	84	279	0	5	18	0
production	130	102	6	9	28	0
wild	96	3	216	35	30	0
...	...	...	...	...	...	...



$$\vec{v}_{water}(plant) = \sum_{w} f(plant, w) * f(water, w) * \vec{e}_w$$

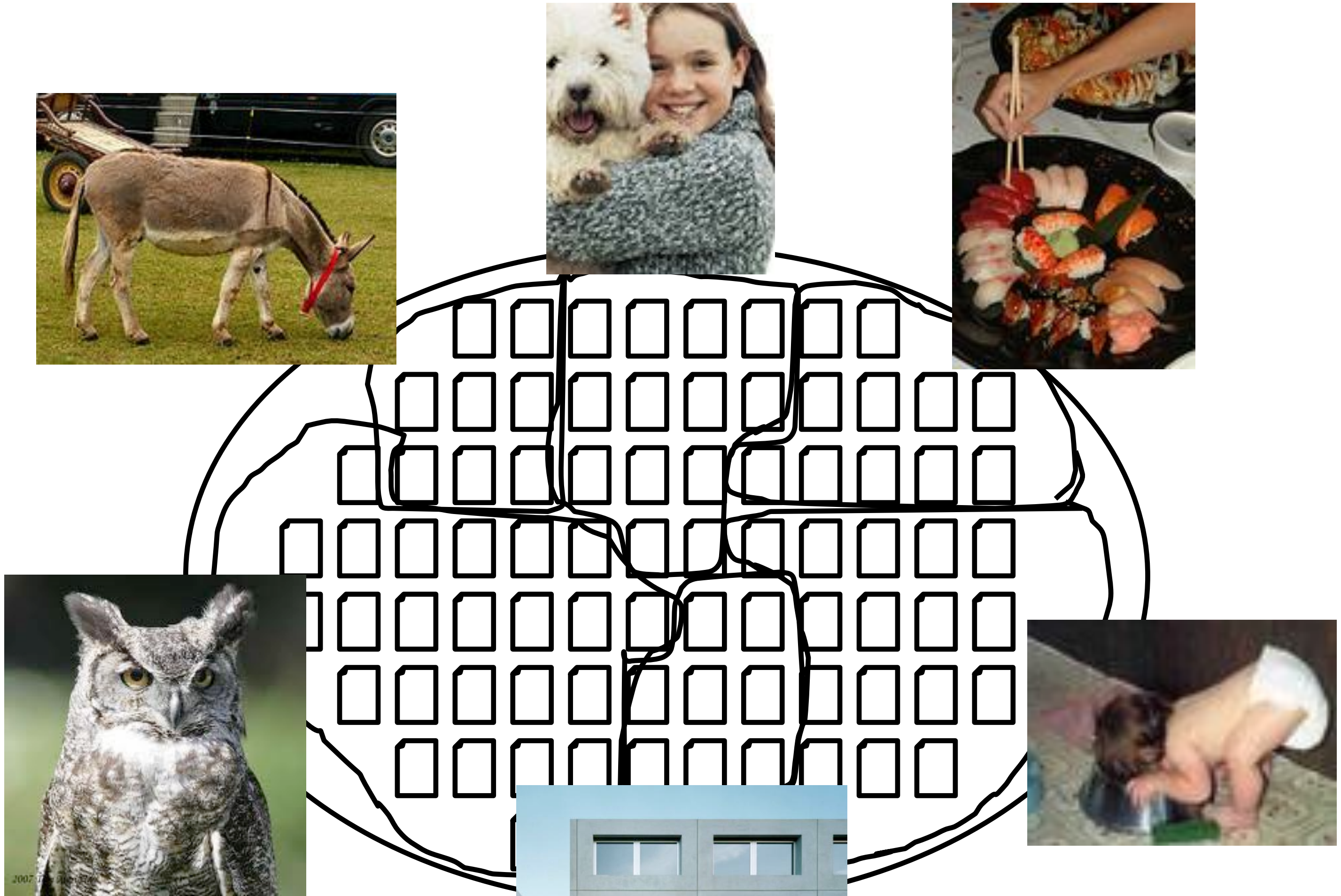
Erk&Padó 2008, Thater et al. 2011

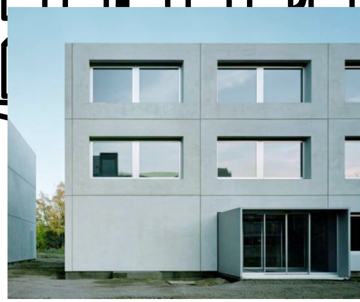
Including Non-Linguistic Context



Titov &Kozhevnikov 2010

Including Non-Linguistic Context





ESP Game Dataset

MS Video Description Corpus

Grounding Distributional Semantics in Visual Information

ESP Game Dataset

MS Video Description Corpus

A Corpus of Cooking Scenes

- 41 basic cooking tasks
- 212 high-resolution video recordings (4-8 videos per task, varying subjects, 4.5 min. on average)

A Cooking Video

Rohrbach et al. 2012, Regneri et al. 2012

Low-Level Annotation for Cooking Scenes

- 41 basic cooking tasks
- 212 high-resolution video recordings (4-8 videos per task, varying subjects, 4.5 min. on average)
- Annotated with activity labels and associated objects

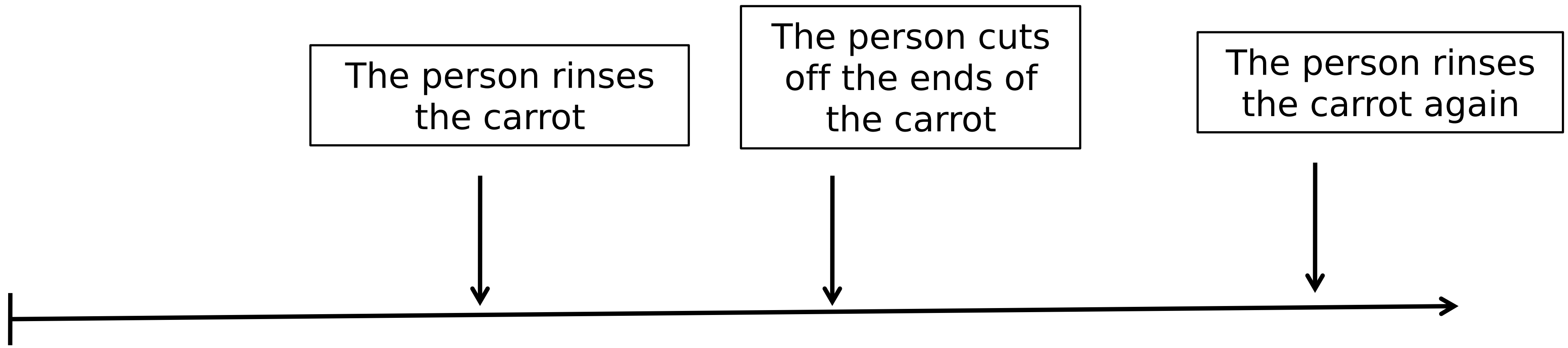
896 -1137	wash	[hand,carrot]
1145 -1212	shake	[hand,carrot]
1330 -1388	close	[hand,drawer]
1431 -1647	take out	[hand,knife,drawer]
1647 -1669	move	[hand,cutting board,counter]
1673 -1705	move	[hand,carrot,bowl,cutting board]
1736 -1818	cut	[knife,carrot,cutting board]
1919 -3395	slice	[knife,carrot,cutting board]

The TACoS Corpus

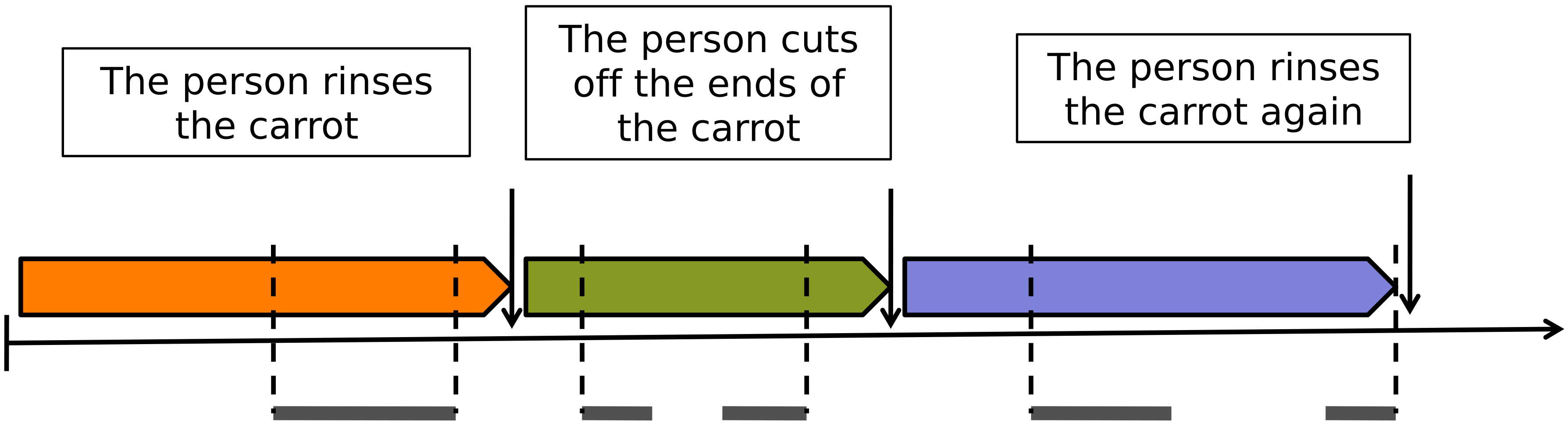
- TACoS: Saarbrücken Corpus of Textually Annotated Cooking Scenes
  - Cooking videos + low-level annotation
  - Multiple (20) natural-language descriptions of each video collected via M-Turk
  - Aligned with video on sentence level
  - Resulting in 17,000 sentence – video segment pairs

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B. & Pinkal, M.: Transactions of ACL 2013

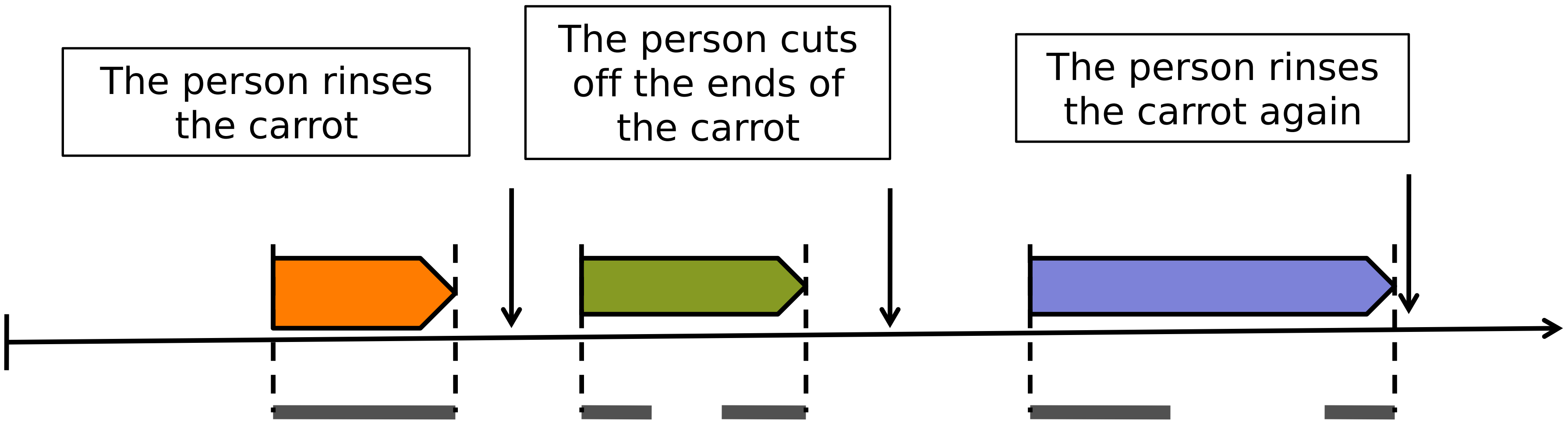
Alignment of Video Descriptions



Alignment of Video Descriptions



Alignment of Video Descriptions



A TACoS Sample

frame	start	end	action	participants	sequence 1	sequence 2	sequence 3
	743	911	wash	hand, carrot	He washed carrot	The person rinses the carrot.	He rinses the carrot from the faucet
	982	1090	cut	knife, carrot, cutting board	He cut off ends of carrots	The person cuts off the ends of the carrot.	He cuts off the two edges.
	1164	1252	open	hand, drawer		The person searches for the trash can, then throws the ends of the carrot away.	He searches for something in the drawer, failed attempt, he throws away the edges in trash.
	1679	1718	close	hand, drawer			
	1746	1799	trash	hand, carrot			He rinses the carrot again.
	1854	2011	wash	hand,carrot	He washed carrot	The person rinses the carrot again.	
	2011	2045	shake	hand,carrot			He starts chopping the carrot into small pieces.

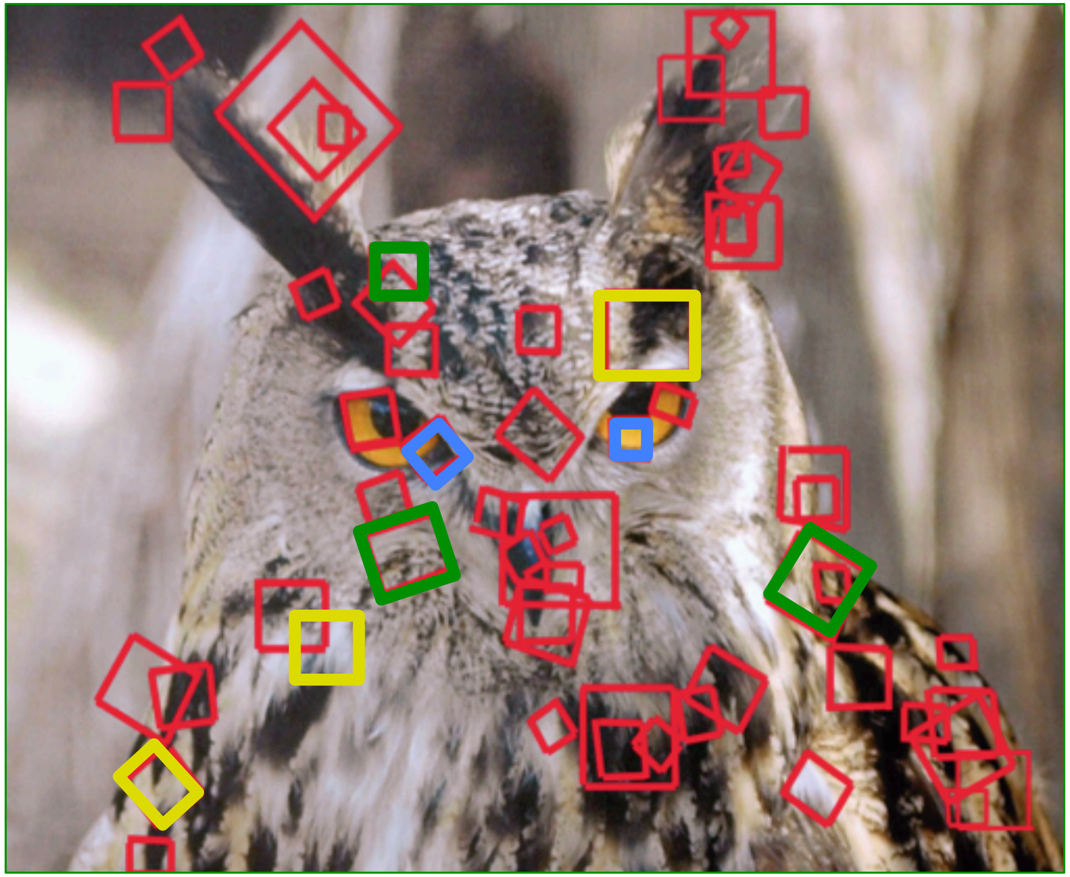
TACoS: Linguistic Variation and Granularity of Action Descriptions

- Variation in lexical realization: 435 verb lemmas vs. 58 low-level activity labels
- Variation in granularity: 2.7 low-level tags covered by one action description, on average

Modeling Similarity of Action Descriptions

- The task: Provide models of distributional similarity that matches human similarity ratings of action descriptions
- The models:
  - Video-based models
  - Text-based models
  - Combinations of the two
- Evaluation on a newly created dataset ("ASim" dataset), consisting of pairs of action descriptions and human similarity ratings.

Visual Words



Feng&Lapata 2010, Bruni et al. 2011

Visual Words in Videos

Rohrbach et al. 2012, Regneri et al. 2012

Distributional Models

- Video-based models
  - BOW vectors (16,000 dimensions)
  - Vectors obtained from visual classifier output
  - Combination of the two
- Text-based models
  - Jaccard coefficient
  - Contextualization model of Thater et al. 2011
  - Combination of the two
- Combination of text- and video-based models
  - by averaging the similarity scores

The Evaluation Dataset

- 900 pairs of action descriptions (TACoS sentences),
  - annotated with similarity scores between 1 and 5 (similarity with respect to "how the action was carried out")
- Sentence pairs either share the object or the verb
  - *The man washes the carrot. – She dices the carrot.*
  - *The man washes the carrot. – A woman washes an apple under the faucet.*
- Sentences describe reasonably frequent activities
  - CUT, SLICE, CHOP, PEEL, TAKE_APART, WASH

Evaluation Results

	Model	Same Object	Same Verb	Overall
Text	Jaccard	0.28	0.25	0.25
	Textual Vectors	0.30	0.25	0.27
	Text Combined	0.39	**0.35**	0.36
Video	Visual Raw Vectors	0.53	-0.08	0.35
	Visual Classifier	0.60	0.03	0.44
	Video Combined	0.61	-0.04	0.44
Mix	All Unsupervised	0.58	0.32	0.48
	All Combined	**0.67**	0.28	**0.55**
Upper Bound		0.84	0.43	0.73

Summary of Results

- First distributional model for action descriptions
- Visual context outperforms textual context
- Combination approaches upper bound of interrater agreement
- ... and there is much space left for improvement

Outlook

- Try more sophisticated methods to combine textual and visual information.
- Use TACoS for the generation of text from videos (Rohrbach et al., submitted).
- Leverage the discourse-level information in TACoS, and combine it with script knowledge to improve grounded models of word meaning, video understanding, and generation.