The Visual Denotations of Sentences

Julia Hockenmaier
with Peter Young and Micah Hodosh
juliahmr@illinois.edu
University of Illinois

Sentence-Based Image Description and Search

Hodosh, Young, Hockenmaier, JAIR 2013.

Task: Image Description

Two boys are playing football.

A little girl is enjoying the swings

A little girl is enjoying the swings

A motorbike is racing around

A boy in a yellow uniform c

An elephant is being wa

Conceptual image descriptions...

... describe
the depicted **entities, events, scenes**

... **only** describe
what can be seen from the image

... may differ in the **amount of detail**

Why not caption **generation**?

For image and language understanding,
the **semantic** question of whether a sentence describes an image or not is fundamental

Natural Language Generation has additional **syntactic** and **pragmatic** aspects that detract from the semantic question

Natural Language Generation
is much harder to **evaluate**

Task: Image Search

Two boys are playing football.

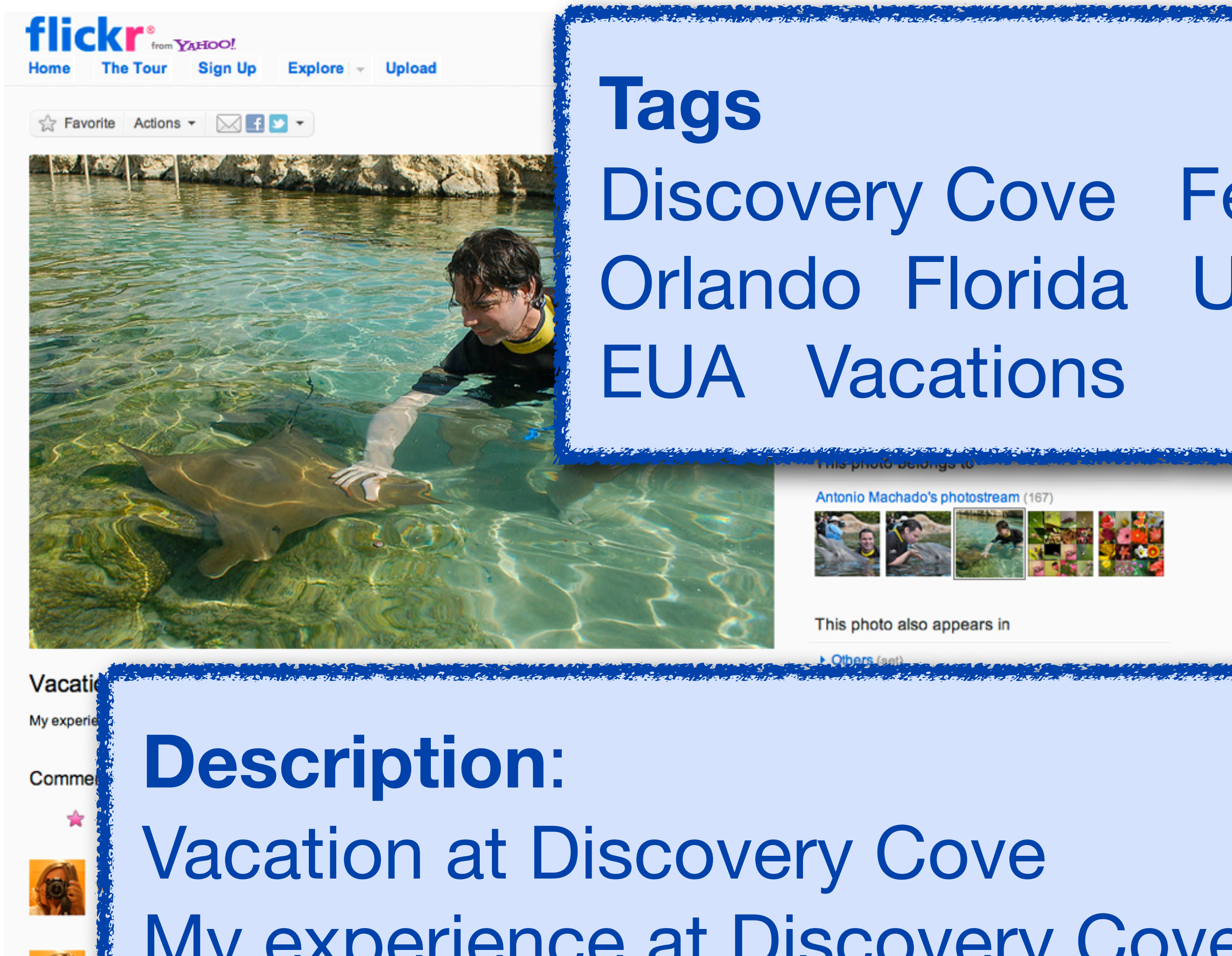
A little girl is enjoying the swings

A little girl is enjoying the swings

A motorbike is racing aroun

A boy in a yellow uniform

An elephant is being wa



Tags

Discovery Cove   Férias

Orlando   Florida   USA

EUA   Vacations

Description:

Vacation at Discovery Cove

My experience at Discovery Cove in Orlando, FL

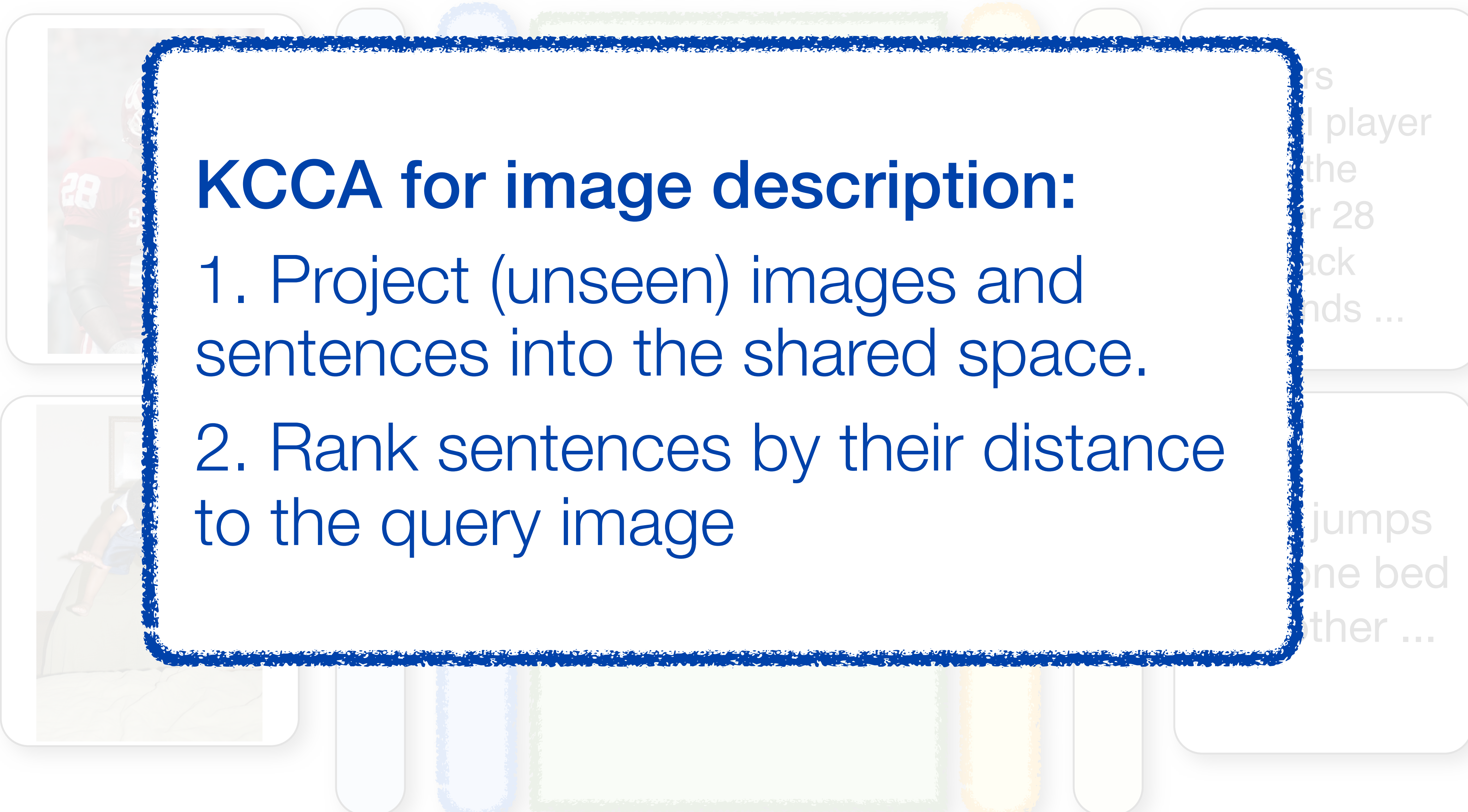
Our data

~32K Flickr images, each annotated with 5 crowdsourced captions.

Our captions

- Four basketball players in action.
- Young men playing basketball in a competition.
- Four men playing basketball, two from each team.
- Two boys in green and white uniforms play basketball with two boys in blue and white uniforms.
- A player from the white and green highschool team dribbles down court defended by a player from the other team.

Our model: Kernel CCA



KCCA for image description:

1. Project (unseen) images and sentences into the shared space.
2. Rank sentences by their distance to the query image

Experimental Results

Rate of success (S@k)

	Image annotation			Image search		
	S@1	**S@5**	**S@10**	**S@1**	**S@5**	**S@10**
NN	5.8***	15.3***	20.1***	4.9***	12.9***	18.1 ***
BoW1	12.2***	30.3***	39.7***	11.4***	30.5***	40.2 ***
BoW5	15.0*	34.1**	42.7***	12.1***	31.5***	40.8 ***
TagRank	16.2	34.2**	42.9***	12.4***	31.5***	41.6 ***
Tri5	16.4	32.9***	43.4***	13.1**	33.1**	43.8 ***
Tri5_{Sem}	**16.6**	**37.7**	**49.1**	**15.7**	**36.9**	**48.5**

S@k: Percentage of test items for which the top k results contain a relevant item

Score: 4 (No errors)

A girl wearing a yellow shirt and sunglasses smiles.

A man climbs up a sheer wall of ice.

Score: 3 (Minor errors)

A boy jumps into the blue pool water.

A child jumping on a tennis court.

Score: 2 (Major errors)

A dog in a grassy field, looking up .

A boy in a blue life jacket jumps into the water .

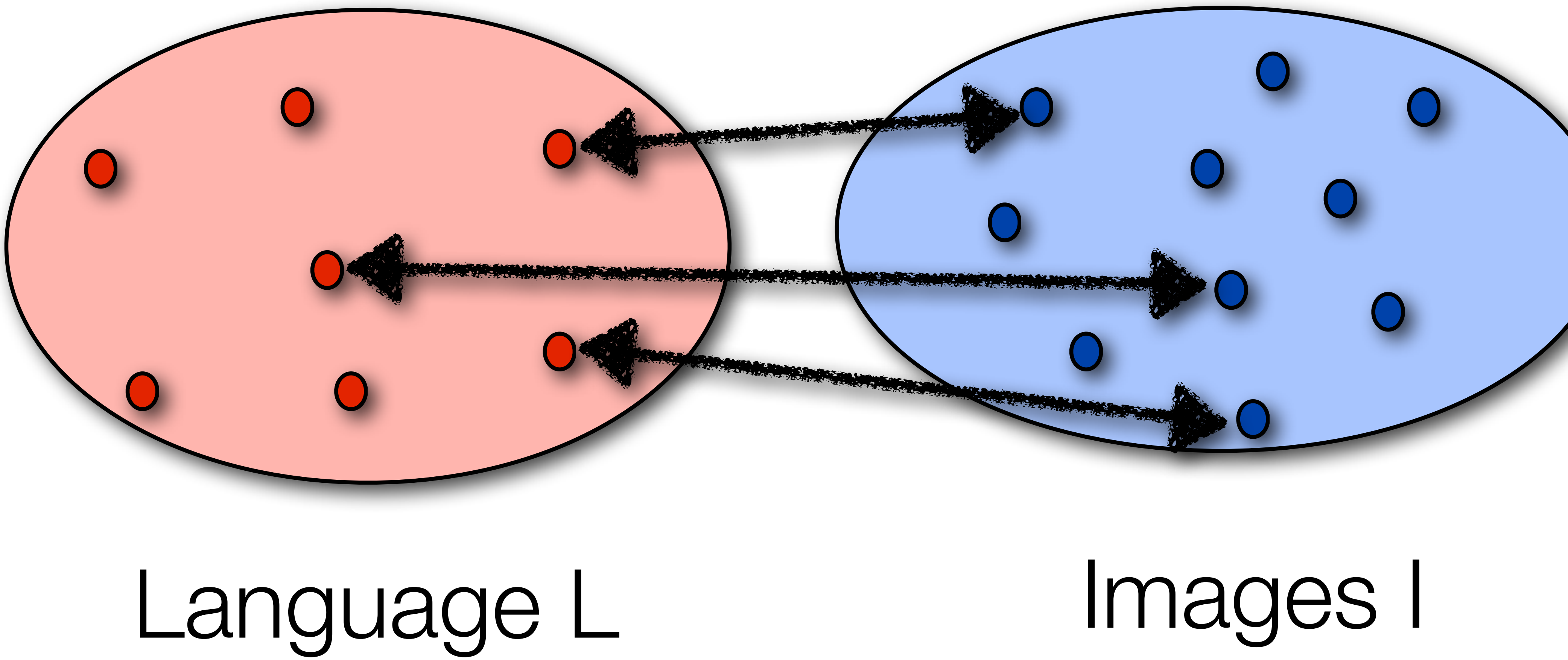
Score: 1 (Unrelated)

Basketball players
in action.

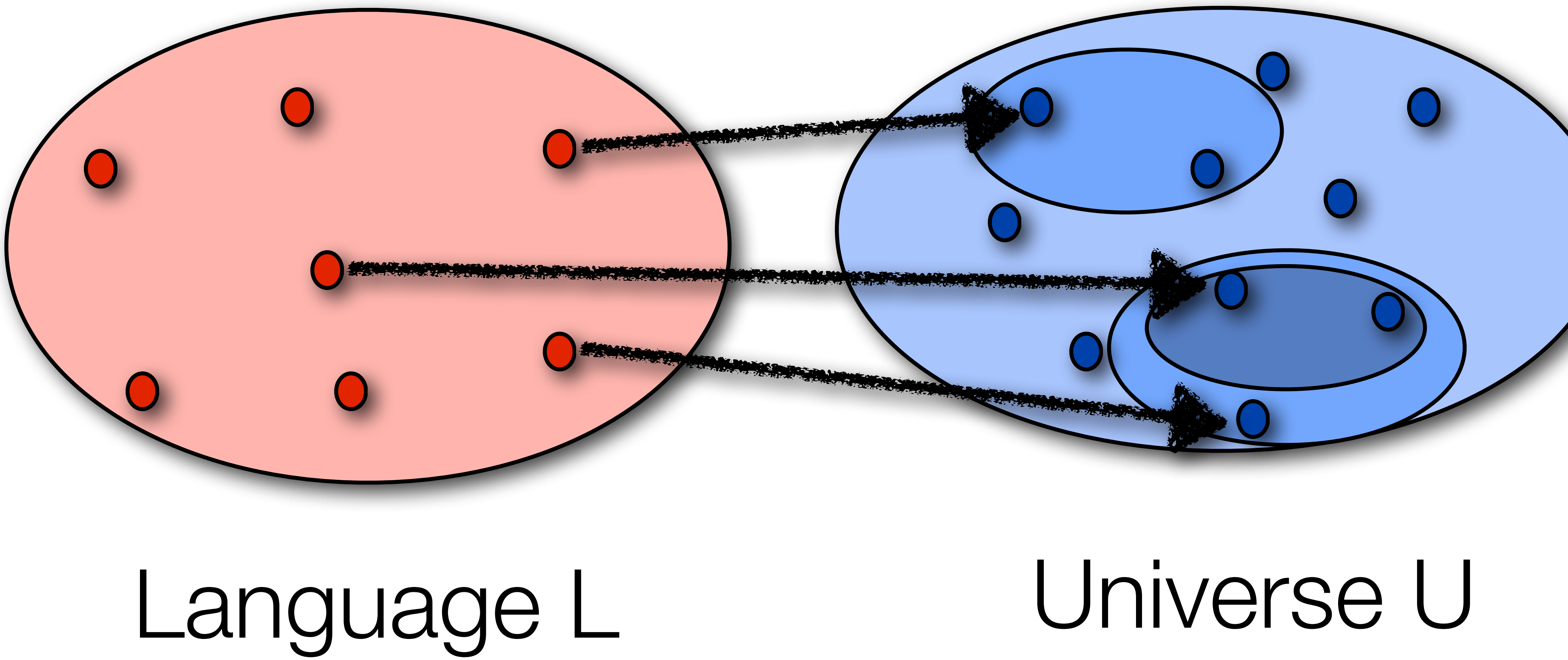
A black dog with a
purple collar running.

Back to semantics…

Implied Semantics



Denotational Semantics



Denotational Semantics

The denotation of a (declarative) sentence is the **set of all possible worlds** in which it is true:

$$[\![s]\!] = \{w \in U: s \text{ is true in } w \}$$

Visual denotations

The visual denotation of a (descriptive) sentence is the **set of all images** for which it is a correct description:

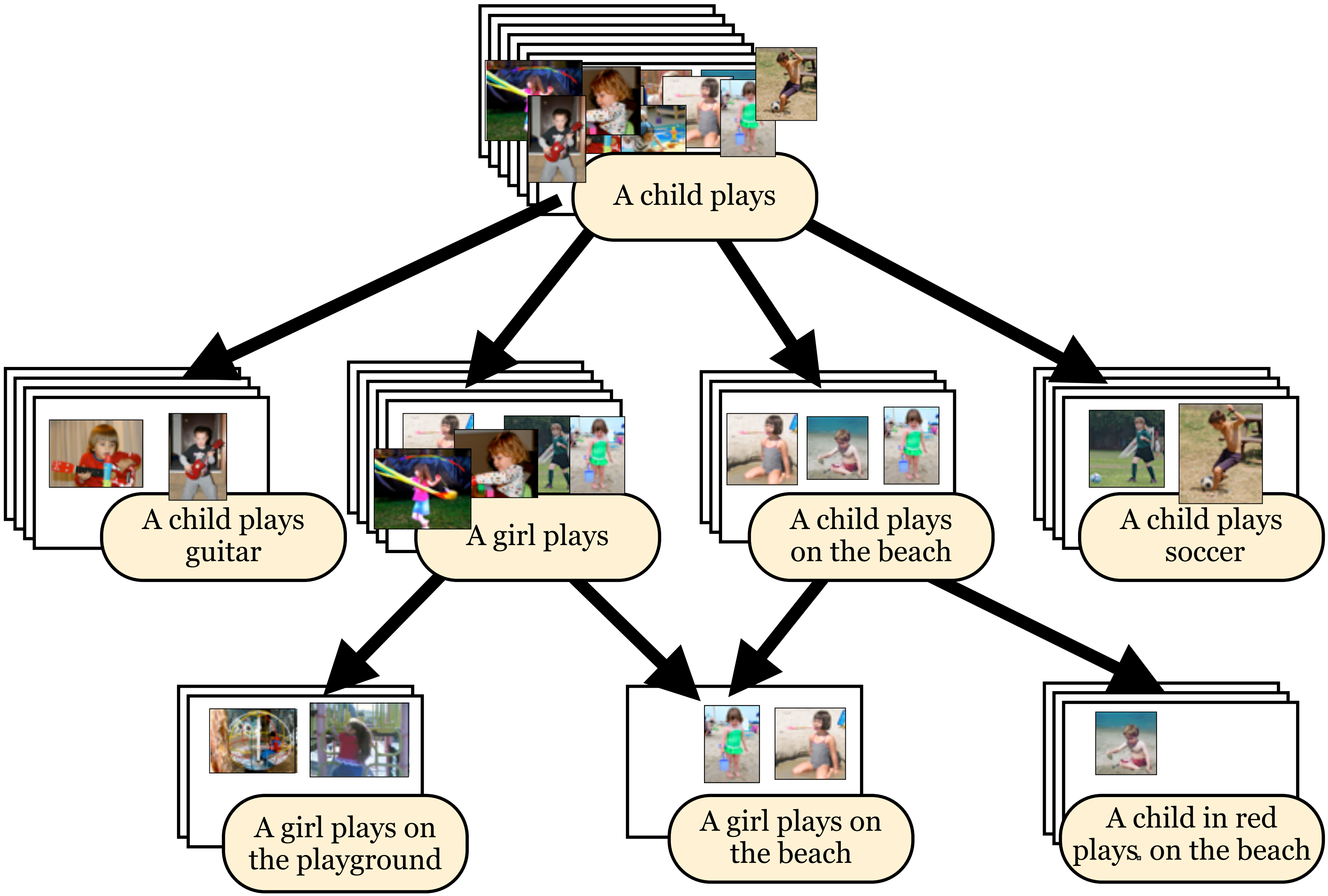
$[\![s]\!] = \{i \in I\text{: } s \text{ describes (part of) } i\ \}$

Denotation Graph

1. Normalize captions:
   - Spelling; capitalization
   - Lemmatization
   - Normalize determiners
2. Make captions more generic:
   - Replace nouns by hypernyms
   - Drop modifiers (adjectives, adverbs, PPs)
3. Extract VPs and NPs

This yields a large subsumption hierarchy
of (partial) image descriptions

Subsumption hierarchy



Statistics

Original data (~32,000 images)

~160K distinct captions

Denotation graph:

~1500K distinct captions:

- ~280K captions with $[\![s]\!] \geq 2$
- ~40K captions with $[\![s]\!] \geq 5$
- ~19K captions with $[\![s]\!] \geq 10$
- ~1.7K captions with $[\![s]\!] \geq 100$
- 142 captions with $[\![s]\!] \geq 1000$
  - e.g. *person play instrument*, *woman standing*, ...

Applications

Better models for
image description/search?

Better models of
natural language semantics?

‘Denotational similarities’

$p(\ \mathrm{VP}_1 \mid \mathrm{VP}_2\ )$

p(talk | engage in conversation) = 0.79

p(play tennis | swing racket) = 0.82

p(stand | wait for subway = 0.58

p(sit | ride subway) = 0.56

p(stand | lean against building) = 0.53

p(shave | look in mirror) = 0.41

p(dig hole | use shovel) = 0.38

p(make face | stick out tongue) = 0.38

Future/Ongoing work

Using denotational similarities:

e.g. for Textual Similarity, Entailment Recognition

Capturing compositionality in our models:

Integrate with (syntactic) grammar induction
for Combinatory Categorial Grammar
(Bisk and Hockenmaier 2012, 2013)

Improving coverage of the denotation graph:

Reduce sparsity of existing captions
Add more images (using other resources?)