Describing Images in Natural Language Part I

EACL tutorial
Julia Hockenmaier
University of Illinois
juliahmr@illinois.edu

Overview

Part 1: High-Level Introduction to Sentence-Based Image Description

- What do we mean by image description?
- What kind of data sets are available?
- What kind of tasks have been proposed?
- How do we evaluate image description systems?
- A proposal for a shared task

Part 2: Digging deeper and going further

- Visual features for image description
- Image description systems
- Image description and semantics

What is Sentence-based Image Description?

How would you describe this image?

A guy in a wetsuit petting a stingray

yes

Somebody kneeling down to touch a really flat fish

perhaps

Does that thing bite?

no

Why would you want to describe images?

Because you get to look at pictures of the cool things people do on their vacations?

Because you get to address some of the most fundamental problems in natural language understanding and artificial intelligence.

Because you get to work on a new, challenging task that could become really important.

Why would you want to describe images?

A test for grounded language understanding:
Image description requires the ability to associate sentences with images that depict the events, entities and scenes they describe.

A test for image understanding/vision:
Image description requires the ability to detect events, entities, and scenes in images.

Why would you want to describe images?

Traditional image retrieval maps text queries to text near the image. But the pictures you get from your camera/phone have no text associated with them.

We'd like to be able to associate text queries *directly* with images.

Sentence-based image description should improve:

... **Image search for everybody**

... **Accessibility to image collections for the visually impaired**

What do we want to say about an image?

How would you describe this image?

A boy in a yellow uniform carrying a football blocks another boy in a blue uniform.

yes

Two boys are playing

perhaps

A dog is running on the beach.

no

How would you describe this image?

Jake tackled Kevin really hard.

perhaps

Last Sunday's game was really rough.

probably not

How would you describe this image?

There's a lot of vibrant blue and some pale green.

probably not

The image shows some shiny surfaces.

probably not

Image descriptions...

... should describe the depicted entities, events, scenes:
Who did what to whom, when and where?

... should *only* describe what is in the image:
No background information that cannot be seen.

... may differ in the *amount of detail* they provide
Each image has many correct descriptions.
Each sentence may describe many different images.

Image descriptions

[Shatford, Jaimes et al., Hollink et al.]

Perceptual image descriptions

- What kind of image?
  (photo vs. drawing, macro, panorama)
- Colors, textures, shapes

Non-visual image descriptions

- Additional context (*Last Sunday's game*)
- Metadata (*Nikon D90, f2.8, GPS coordinates*)

Image descriptions

[Shatford, Jaimes et al., Hollink et al.]

Conceptual image descriptions:
Who did what where to whom?

What events, scenes, entities are depicted?

- Generic: *Kids playing football.*
- Specific: *Jake tackling Kevin.*
- Abstract: *Childhood; Competition*

Most appropriate for image search etc., and for image description as a test for language understanding.

Summary:
What is image description?

Definition of sentence-based image description:

Sentence-based image description is the task of associating images with natural language sentences that describe what entities, events and scenes are depicted in them.

Applications of sentence-based image description:

- Searching online or personal image collections
- A testbed for image understanding
- A testbed for grounded language understanding

Data Sets for

Image Description

Data sets for sentence-based image description

To develop and evaluate sentence-based image description systems, we need corpora of images paired with appropriate captions.

- What data sets are available?
- What strengths and weaknesses do they have?
- What other data could be leveraged for this task?

Data sets for sentence-based image description

Using captioned images from the web (news, photo-sharing sites)

Advantage: Size, 'natural' captions

Disadvantage: Online captions may not describe images

SBU Captioned Photo data set; BBC data set

Using images with purposely created captions

Advantage: Sentences describe the images

Disadvantage: Smaller size, 'unnatural'

IAPR-TC; Illinois Pascal data set, Flickr 8K, Flickr30

News sites often use images just to embellish their stories

Drinking over recommended limit 'raises cancer risk'

COMMENTS (349)

Drinking more than a pint of beer a day can substantially increase the risk of some cancers, research suggests.

The Europe-wide study of 363,988 people reported in the British Medical Journal found one in 10 of all cancers in men and one in 33 in women were caused by past or current alcohol intake.

More than 18% of alcohol-related cancers in men and about 4% in women were linked to

Many people do not know that drinking alcohol can increase their cancer risk.

BBC data set

Police confirmed Tony Blair was inside at the time of the incident

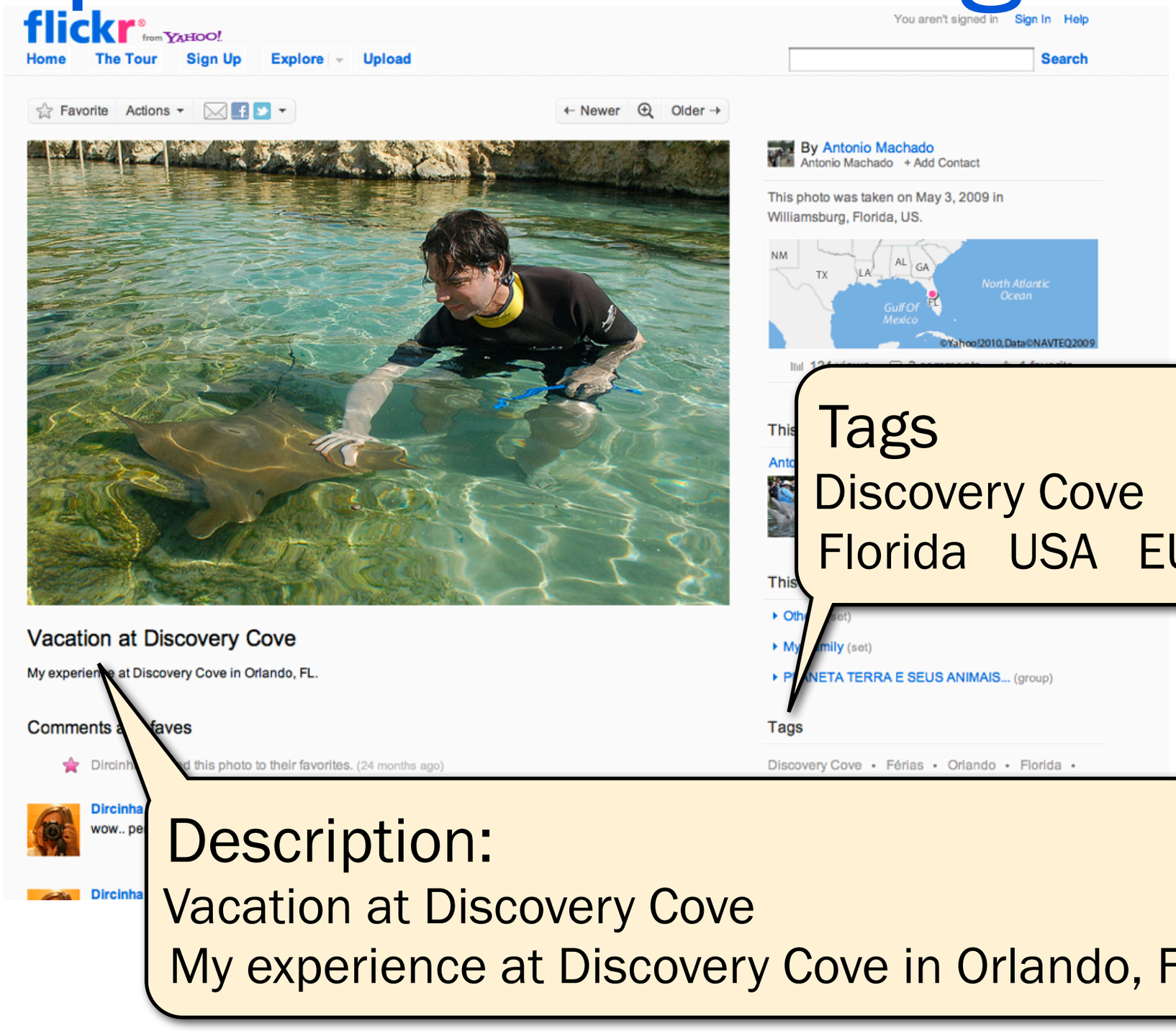
A man has been charged with possessing a knife and assaulting a police officer during an intrusion into the secure area of the prime minister's residence.
Byung Jin Lee, aged 32, was detained after scaling six-foot-high iron railings at Downing Street on Sunday.
Tony Blair was at home during the incident at the back of Number 10, but police said he was not "at risk".
Mr Lee was arrested after a brief struggle and will appear before magistrates on Tuesday.
Scotland Yard said: "We are satisfied that at no time was the prime minister at risk."
Mr Lee is due to appear at the City of Westminster Magistrates Court.
How the drama unfolded outside Downing Street

Crowds flocked to hear techno DJs play

Hundreds of thousands of revellers turned out for the return of Berlin's Love Parade to enjoy a sunshine-filled day of techno music.
The parade returned after a two-year gap caused by financial problems.
Organisers had hoped to attract one million people to the area around the Brandenburg Gate, which was recently home to the World Cup's Fan Mile.
About 40 decorated floats drove through the streets with DJs aboard to entertain the crowds.
The Love Parade started out as a small rave back in 1989 when the Berlin Wall fell.
It continued to grow to a peak of 1.5 million in 1999, but numbers then began falling with commercialism and growing costs blamed for its decline.
It was then scrapped for two years until a German fitness company stepped in with sponsorship in a bid to revive it.
This year, under the banner The Love is back, dance music fans from across the world flocked back to the German capital.
"This is great! I've been waiting three years for this," said Berliner Nicole Koehler, 25. "Hopefully it will be here every year from now on."
The festival went on all day, with parties continuing at the city's nightclubs.
Among those performing were international DJs Paul van Dyk, Westbam and Tiesto

Feng and Lapata 2010

On photo-sharing sites, people describe images...



… but they don't provide conceptual descriptions…

… because they write for (other) people—who can see what's in the picture.

Why bore them?

Gricean maxims:

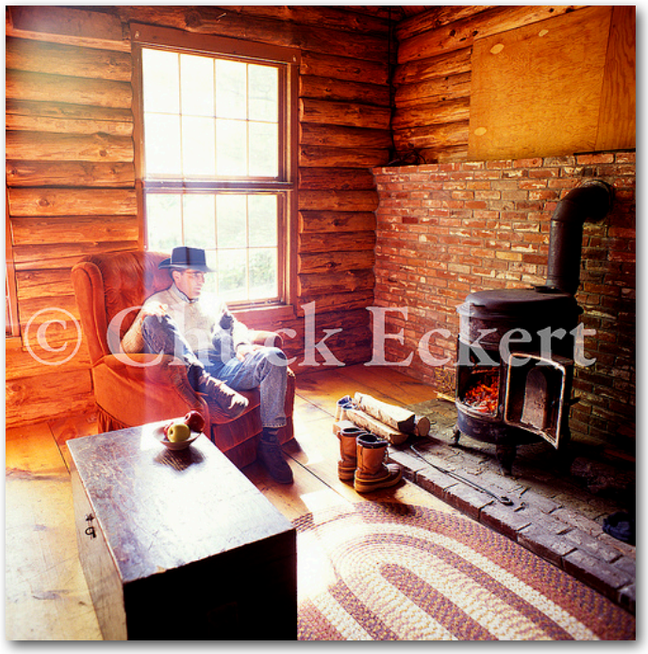
Be informative!

Be relevant!

SBU Captioned Photo Dataset

Not the best idea to roll around on my floor wearing white

Asbestos is used willy-nilly in ukraine as building material. Dad and Tanya have an asbestos roof and an asbestos fence.



Man in mountain cabin sitting by fireplace on a cold winters day

1M images and captions harvested from Flickr

Ordonez et al. 2011

IAPR-TC12 data set

Horse-Riding at the pampas

six people are riding on brown and white horses in a green, flat meadow in the foreground; cows behind them; white and grey clouds in a light blue sky in the background;

Buenos Aires, Argentina
9 December 2004

Panoramic View of the Iguazu Waterfalls

a cascading waterfall in the middle of the jungle; front view with pool of dirty water in the foreground;
this picture was taken from the Brazilian side;

Foz do Iguaçu, Brazil
March 2002

20,000 manually annotated and segmented images
Grubinger et al. 2006; Escalante et al. 2010

Illinois PASCAL data set

A grounded passage plane in a terminal.
An Air Pacific airplane sitting on the tarmac.
Large white commercial airliner parked on runway.
The back and right side of a parked passenger jet.
The passenger plane is sitting at the airport.

A hand holding bird seed and a small bird.
A person holding a small bluebird.
A person holds a bird and seeds.
A small bird is sitting on a person's hand that has bird seed in it.
A small black, white, and brown bird perched on and eating out of a man's hand.

1,000 images from the PASCAL VOC 2008 challenge (20 object categories) with 5 crowdsourced captions
Rashtchian et al. 2010

Illinois Flickr8k/30k data sets

A goalie in a hockey game dives to catch a puck as the opposing team charges towards the goal.
The white team hits the puck, but the goalie from the purple team makes the save.
Picture of hockey team while goal is being scored.
Two teams of hockey players playing a game.
A hockey game is going on.

A group of people are getting fountain drinks at a convenience store.
Several adults are filling their cups and a drink machine.
Two guys getting a drink at a store counter.
Two boys in front of a soda machine.
People get their slushies.

32k images of people (and dogs) from Flickr with 5 crowdsourced captions

Rashtchian et al. 2010, Hodosh et al. 2013, Young et al. 2014

Image description with Amazon Mechanical Turk

Instructions:

Describe the objects and actions; Use adjectives; be brief

5 captions per image

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.

A man crouched on a snowy peak.
A man in a green jacket stands in deep snow
at the base of a mountain.
A man kneels in the snow.
A man measures the depth of snow.
A mountain hiker is digging stakes
into the thick snow.

Image Description Tasks

Task definitions

Generate captions directly from image features:

- Requires an explicit mapping between image & text.
- Requires a surface realization model to guarantee fluency etc.
- Requires human evaluation of correctness & grammaticality.

Transfer (and combine) captions from similar images:

- Requires a unimodal (image) similarity metric
- May also require a surface realization model.
- Requires human evaluation of correctness & grammaticality.

Score and rank a pool of captions for each image:

- Requires a cross-modal (image-sentence) similarity metric.
- Benefits from human relevance judgments,
- but may be evaluated automatically.

The underlying semantic task

Learn an affinity function f(**i**, **s**) over images **i** and sentences **s** that is maximized when **s** describes **i**.

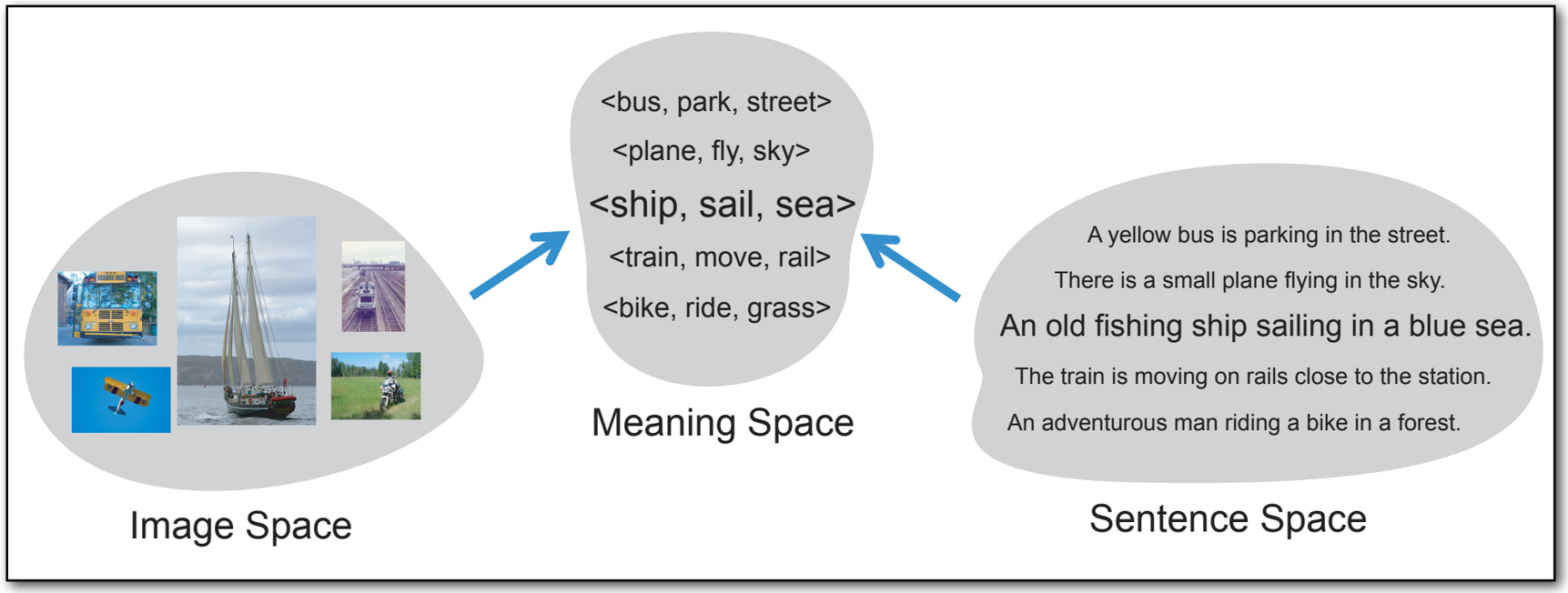
Image search:

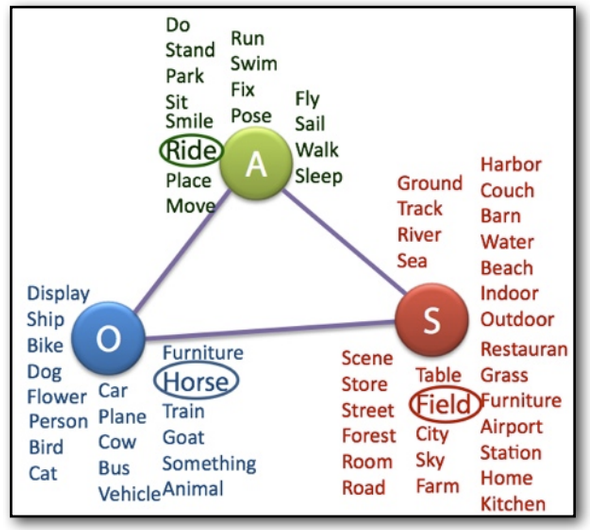
$\mathbf{i}^* = \text{argmax}_{\mathbf{i} \in \mathbf{I}_{cand}} f(\mathbf{i}, \mathbf{s}_{query})$

Image annotation:

$\mathbf{s}^* = \text{argmax}_{\mathbf{s} \in \mathbf{S}_{cand}} f(\mathbf{i}_{query}, \mathbf{s})$

Mapping images & sentences to an explicit semantic space





Farhadi et al. 2010

Semantics of images/sentences:≈ ⟨Object, Action, Scene⟩

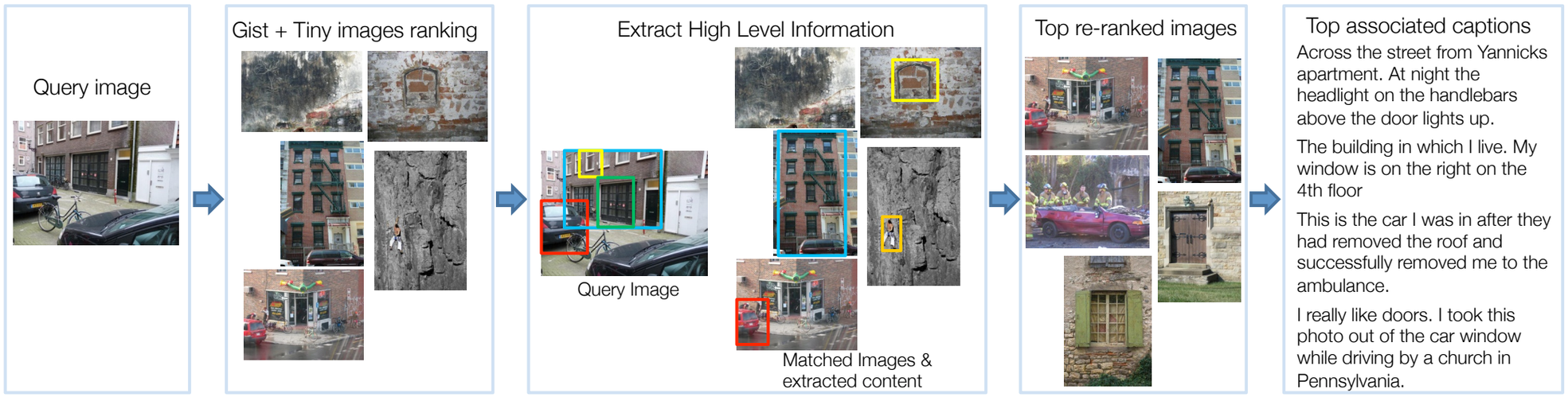
Use Markov Random Fields to predict most likely meaning triplet for images and sentences.

Challenges for explicit semantic mappings

Scalability:

- Restricted to predefined sets of semantic classes (objects, actions, scenes)
- Requires accurate detectors for each semantic class
  - Performance on Illinois PASCAL data may not be indicative of performance on other data (e.g. Flickr images)

Image description as retrieval



Ordonez et al. 2011:
Find most similar image among 1M Flickr images (SBU data set), and re-rank their captions.

Advantages of retrieval

Unimodal image-image similarities are much easier to define than cross-modal image-text similarities.

Can leverage large amounts of human-written captions.

Can be used to provide suggestions for generation-based approaches.

Challenges for retrieval

Ordonez et al.'s approach works because the SBU data set is very large (1 million images):

Query Image | 1k matches | 10k matches | 100k matches | 1million matches

This can only be done on harvested captions, which may not actually describe the images.

Image description as generation

Kulkarni et al.:
This is a picture of three persons, one bottle and one diningtable. The first rusty person is beside the second person. The second person is by the third rusty person. The colorful diningtable is near the first rusty person, and near the second person, and near the third rusty person.
Yang et al.:
Three people are showing th bottle on the street.
Mitchell et al.:
people with a bottle at the table

Kulkarni et al.:
This is a picture of two pottedplants, one dog and one person. The black dog is by the black person, and near the second feathered pottedplant.
Yang et al.:
The person is sitting in the chair in the room.
Mitchell et al.:
A person in black with a black dog by potted plants.

Comparison from Mitchell et al. (EACL 2012)

Advantages of generation

Ultimately, this is the task we need to solve:
Given an image, produce a (possibly novel) sentence that describes the image well.

Easier to define the scope of what is described explicitly:

- Spatial relations between entities?
- Activities?
- Scenes?
- Attributes?

Can leverage other text resources.

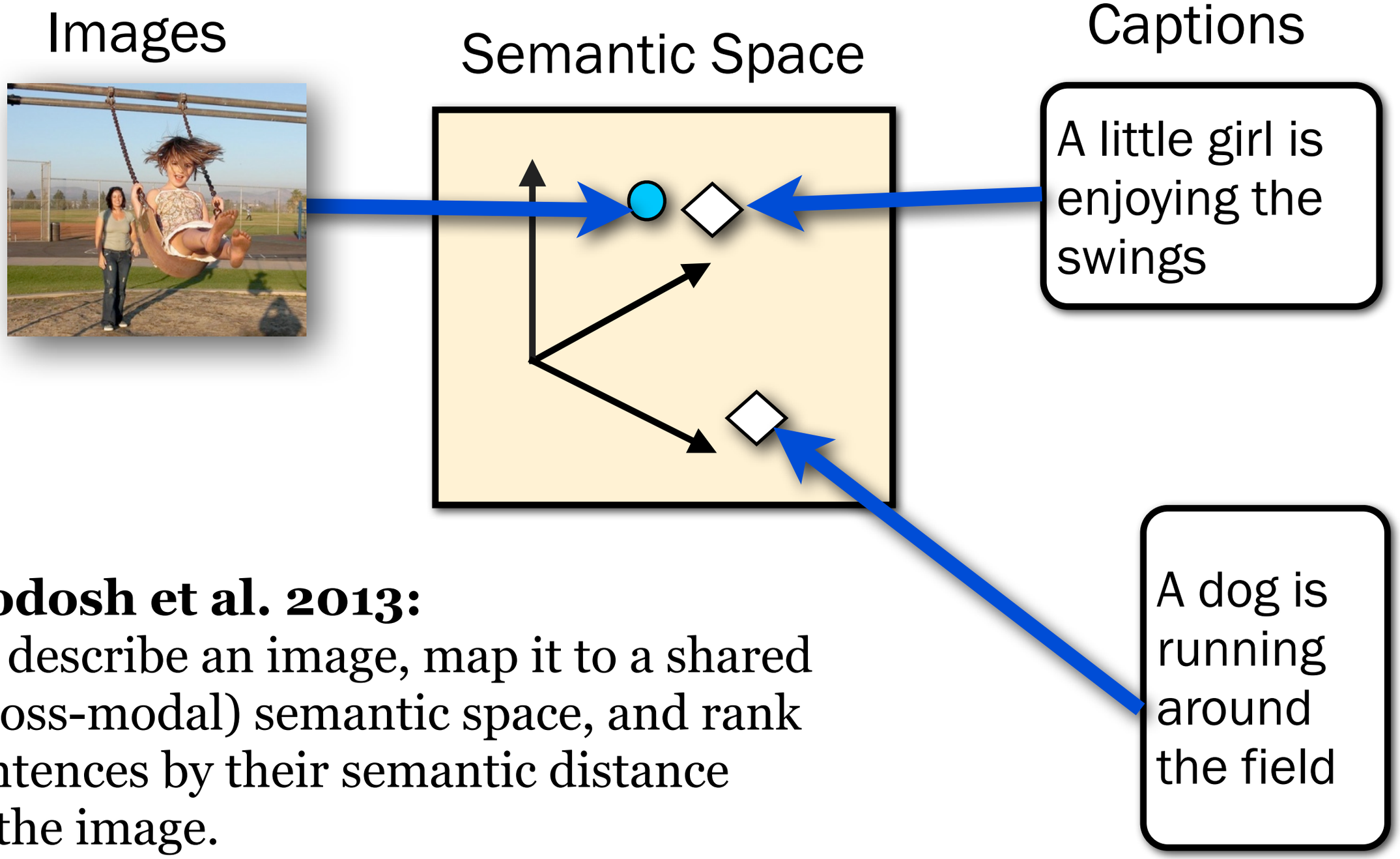
Issues for generation

For image and language understanding, the (truth-conditional) 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:
-Is the caption written in fluent/grammatical English?
-Is this a sentence people would use to describe images?

Natural Language Generation is difficult to evaluate

Image description as cross-modal ranking



Hodosh et al. 2013:
To describe an image, map it to a shared (cross-modal) semantic space, and rank sentences by their semantic distance to the image.

Cross-modal image annotation

A little girl is enjoying the swings

Two boys

People in a line holding lit roman candles.

A little girl is enjoying the swings

A motorbike is racing around a track.

An elephant is being washed.

Cross-modal image search

Two

A little girl is enjoying the swings

People in a line holding lit roman candles.

A little girl is enjoying the swings

A motorbike is racing around a track.

An elephant is being washed.

Advantages of cross-modal ranking

Aims to directly measure the quality of the semantic mapping between images and sentences.

- No need to worry about syntax or pragmatics, as in generation.

Makes it possible to define objective benchmarks that allow different systems to be compared directly.

- Everybody needs to rank the same set of sentences/images.

Can be evaluated automatically.

- Automatic evaluations correlate highly with human judgments.

Natural framework for annotation *and* search.

Issues for cross-modal ranking

Requires large(ish) pool of images paired with human-written sentences.

Benefits from additional human relevance judgments.

Is performance an artifact of what sentences are available in the pool?

Evaluating Image Description

Desiderata for evaluation

Reliability:

- Higher-scoring systems produce better descriptions.
- Human evaluation: high inter-annotator agreement.

Reproducibility and objectivity:

- Repeated evaluation of the same system produces the same scores.
- Different systems can be evaluated independently, and their scores can still be compared directly.

Cost and efficiency:

- Evaluation should be as inexpensive as possible.
- Evaluation should be possible on a large scale.

Evaluating generated image descriptions

Human judgments

- Advantages: Accuracy
- Challenges: Reliability, reproducibility, cost

Automatic comparison to reference captions (BLEU, ROUGE)

- Advantages: Can be done on a large scale
- Challenges: Poor correlation with human judgments

Evaluating ranking-based description systems

Automatic evaluation is possible:
One image in the test pool is the image that the query sentence was written for.

- What is the median rank of that image?
- Recall@K: How often do we retrieve this image as the first result, among the first 5 results, among the first 10 results?

Challenges for evaluating ranking-based systems

The test pool may contain other images for which the test query is a good description.
- Automatic evaluation underestimates performance

We need to collect human judgments for all image-sentence pairs in the test pool.
- Prohibitively expensive to do this exhaustively, but:
- BLEU is a useful filter to identify implausible pairings
- Unlike for NLG, the same judgments can be re-used in evaluating different systems.

Collecting human relevance judgments

Binary task:

Which of these 10 captions describes the image, possibly with minor errors

Crowdsourcing:

On a test set of 1,000 image-caption pairs, we collected 3 judgments for the 10 highest results of 16 models on annotation and search

Preprocessing

Use BLEU (with stemming and stop word removal) to filter out implausible image-caption pairs.
Removes 86% of all pairs, but only 6.7% of good pairs

Using human relevance judgments

Rate of success (S@k):

What percentage of queries has at least one relevant response among the top k results?

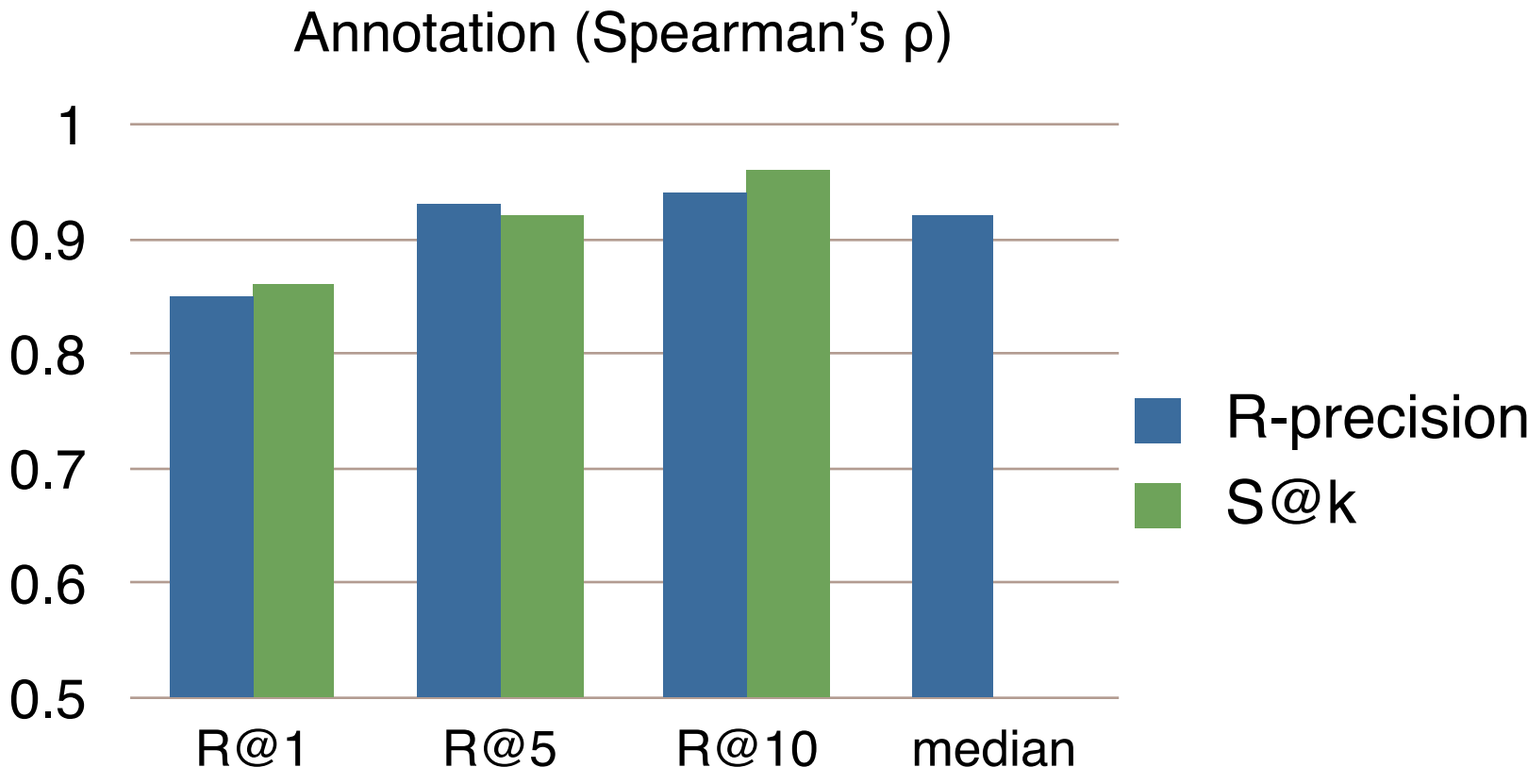
R-precision:

If query q_i has r_i relevant responses, report precision at rank r_i for q_i.

Standard IR metric when the number of relevant responses varies by query.

Can evaluation be automated?

Do automated metrics rank different (n = 30) systems the same way as human metrics?



Possibly, if you go beyond the first result

A proposal for a shared task

Why a shared task?

We need objective evaluation metrics and benchmarks to compare different systems directly.

(cf. CoNLL shared tasks, PASCAL VOC challenges, ImageNet)

We need reproducible experiments and evaluations.

No repeated human evaluation should be required.

We need data sets that both vision and language people can use.

We want to lower the barrier of entry for people from both fields.

Ranking-based image description

Test data:

A large pool of unseen images and their captions

Image annotation task:

For each test image, rank the test captions

Image search task:

For each test caption, rank the test images

Evaluation:

Without relevance judgments: median rank of gold item, R@k
With (exhaustive) relevance judgments: R-precision, S@k

Data sets

Hodosh et al. 2013 (JAIR):
Illinois Flickr 8k as benchmark for ranking-based image annotation & search

- Training set: 6,000 Flickr images with 5 crowdsourced captions
- Test set: 1,000 Flickr images with 1 crowdsourced caption each (+human relevance judgments)

Illinois Flickr 30k data set (Young et al., TACL, 2014):

- Mostly Creative-Commons licensed images

Alternative: IAPR-TC12

- (20,000 segmented images with captions)

REFERENCES

References

Image description in general
Jaimes, A., Jaimes, R., & Chang, S.-F. (2000). A conceptual framework for indexing visual information at multiple levels. In Internet Imaging 2000, Vol. 3964 of Proceedings of SPIE, pp. 2–15, San Jose, CA, USA.
Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. Cataloging & Classification Quarterly, 6, 39–62.
Data sets (Images & sentences)
Everingham, M., Gool, L. V., Williams, C., Winn, J., & Zisserman, A. (2008). The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. http://www.pascal-network.org/challenges/VOC/voc2008/workshop/.
Grubinger, M., Clough, P., Müller, H., & Deselaers, T. (2006). The IAPR benchmark: A new evaluation resource for visual information systems. In OntoImage 2006, LREC Workshop, pp. 13–23.
Hugo Jair Escalante, Carlos Hernandez, Jesus A. Gonzalez, Aurelio Lopez, Manuel Montes, Eduardo Morales, Enrique Sucar, Luis Villasenor, and Michael Grubinger (2010) The Segmented and Annotated IAPR-TC12 Benchmark. Computer Vision and Image Understanding Journal, 114(4):419–428, 2010.
Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010). Collecting image annotations using Amazon's Mechanical Turk. In NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk, pp. 139–147.
Feng, Y., & Lapata, M. (2010). How many words is a picture worth? Automatic caption generation for news images. ACL, pp. 1239–1249.
Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. NIPS 24, pp. 1143–1151.
Hodosh, M., Young, P. & Hockenmaier J. (2013). Framing image description as a ranking task: data, models, and evaluation metrics, Journal of Artificial Intelligence Research, Volume 47, pages 853-899
Young, P., Lai, A., Hodosh M & Julia Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL 2014.

References

Sentence-based image description

Feng, Y., & Lapata, M. (2008). Automatic image annotation using auxiliary text information ACL-08: HLT, pp. 272–280.

Feng, Y., & Lapata, M. (2010). How many words is a picture worth? Automatic caption generation for news images. ACL, pp. 1239–1249.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. ECCV Part IV, pp. 15–29

Kuznetsova, P., Ordonez, V., Berg, A., Berg, T., & Choi, Y. (2012). Collective generation of natural image descriptions. 50th ACL (Volume 1: Long Papers), pp. 359–368.

Li, S., Kulkarni, G., Berg, T. L., Berg, A. C., & Choi, Y. (2011). Composing simple image descriptions using web-scale n-grams. CoNLL, pp. 220–228.

Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., & Berg, T. L. (2011). Baby talk: Understanding and generating simple image descriptions. CVPR, pp. 1601–1608.

Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A., Berg, T., & Daume III, H. (2012). Midge: Generating image descriptions from computer vision detections. 13th EACL, pp. 747–756.

Yang, Y., Teo, C., Daume III, H., & Aloimonos, Y. (2011). Corpus-guided sentence generation of natural images. EMNLP, pp. 444–454

Gupta, A., Verma, Y., & Jawahar, C. (2012). Choosing linguistics over vision to describe images. AAAI.

Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. NIPS 24, pp. 1143–1151.

Elliott, D. and Keller F. 2013. Image Description Generation from Structured Image Representations. EMNLP, pp. 1292-1302.

Hodosh, M., Young, P. & Hockenmaier J. (2013). Framing image description as a ranking task: data, models, and evaluation metrics, Journal of Artificial Intelligence Research, Volume 47, pages 853-899