CSC2539 - Datasets and Metrics for Image Caption Generation

Kaustav Kundu

University of Toronto

Types of Image Descriptions

- Conceptual
  - Specific: *Identifying people and locations*
  - Generic: *Related to scene understanding*

Types of Image Descriptions

- Conceptual
  - Specific: *Identifying people and locations*
  - Generic: *Related to scene understanding*
- Non Visual

I don't chew up the couch and pee in the kitchen mama!

Source: *SBU caption dataset*

Patriots quarterback Tom Brady holds the Super Bowl MVP trophy for the third time during a news conference held a day after New England beat Seattle 28-24 in Glendale, Ariz. Brady won a pickup truck for his standout performance and is hoping to give it to teammate Malcolm Butler, whose late-game interception sealed the victory. (Jamie Squire/Getty Images)

Source: *CBC News Website*

Types of Image Descriptions

- Conceptual
  - Specific: *Identifying people and locations*
  - Generic: *Related to scene understanding*
- Non Visual

I don't chew up the couch and pee in the kitchen mama!

Source: *SBU caption dataset*

Patriots quarterback Tom Brady holds the Super Bowl MVP trophy for the third time during a news conference held a day after New England beat Seattle 28-24 in Glendale, Ariz. Brady won a pickup truck for his standout performance and is hoping to give it to teammate Malcolm Butler, whose late-game interception sealed the victory. (Jamie Squire/Getty Images)

Source: *CBC News Website*

- Perceptual

  From a professional photographer's point of view

Types of Image Descriptions

- Conceptual
  - Specific: *Identifying people and locations*
  - Generic: *Related to scene understanding*

  Focus of the today's topic
- Non Visual

I don't chew up the couch and pee in the kitchen mama!

Source: *SBU caption dataset*

Patriots quarterback Tom Brady holds the Super Bowl MVP trophy for the third time during a news conference held a day after New England beat Seattle 28-24 in Glendale, Ariz. Brady won a pickup truck for his standout performance and is hoping to give it to teammate Malcolm Butler, whose late-game interception sealed the victory. (Jamie Squire/Getty Images)

Source: *CBC News Website*

- Perceptual

  From a professional photographer's point of view

Overview

- Datasets for image caption generation
  - Single sentence generation
  - Multiple sentence/paragraph generation

Overview

- Datasets for image caption generation
  - Single sentence generation
  - Multiple sentence/paragraph generation
- Datasets for video caption generation

Overview

- Datasets for image caption generation
  - Single sentence generation
  - Multiple sentence/paragraph generation
- Datasets for video caption generation
- Datasets for referring expressions task

Overview

- Datasets for image caption generation
  - Single sentence generation
  - Multiple sentence/paragraph generation
- Datasets for video caption generation
- Datasets for referring expressions task
- Metrics
  - Image measures
  - Text measures
    - Automatic measures
    - Human based measures

UIUC Pascal Sentence[1]

- A camouflaged plane sitting on the green grass.
- A plane painted in camouflage in a grassy field
- A small camouflaged airplane parked in the grass.
- Camouflage airplane sitting on grassy field.
- Parked camouflage high wing aircraft.

- 1000 images randomly sampled from PASCAL VOC 2008 training + validation data with 20 object categories.
- 5 generic conceptual descriptions per image.

[1] Rashtchian et. al., *Collecting Image Annotations Using Amazon's Mechanical Turk*, 2010. [Dataset Link]

UIUC Pascal Sentence[1]

- A camouflaged plane sitting on the green grass.
- A plane painted in camouflage in a grassy field
- A small camouflaged airplane parked in the grass.
- Camouflage airplane sitting on grassy field.
- Parked camouflage high wing aircraft.

Issues:

- Only 1000 images to train and test models.
- Simple captions and images.
- 25% captions do not contain verbs. 15% contain static verbs like *sit, stand, wear, look*.

[1] Rashtchian et. al., *Collecting Image Annotations Using Amazon's Mechanical Turk*, 2010. [Dataset Link]

Flickr 8k, Flickr 30k

- A biker in red rides in the countryside.
- A biker on a dirt path.
- A person rides a bike off the top of a hill and is airborne.
- A person riding a bmx bike on a dirt course.
- The person on the bicycle is wearing red.

- 8k images in Flickr8k,[2] >30k images in Flickr30k,[3] with 5 descriptions per image.
- More image sentence pairs to train and test models.
- 21% images (vs 40% images in UIUC Pascal Sentence dataset) have static verbs like *sit, stand, wear, look* or no verbs.

---

[2]Hodosh et. al., *Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics*, 2013. [Datset Link]

[3]Young et. al., *From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions*, 2014. [Datset Link]

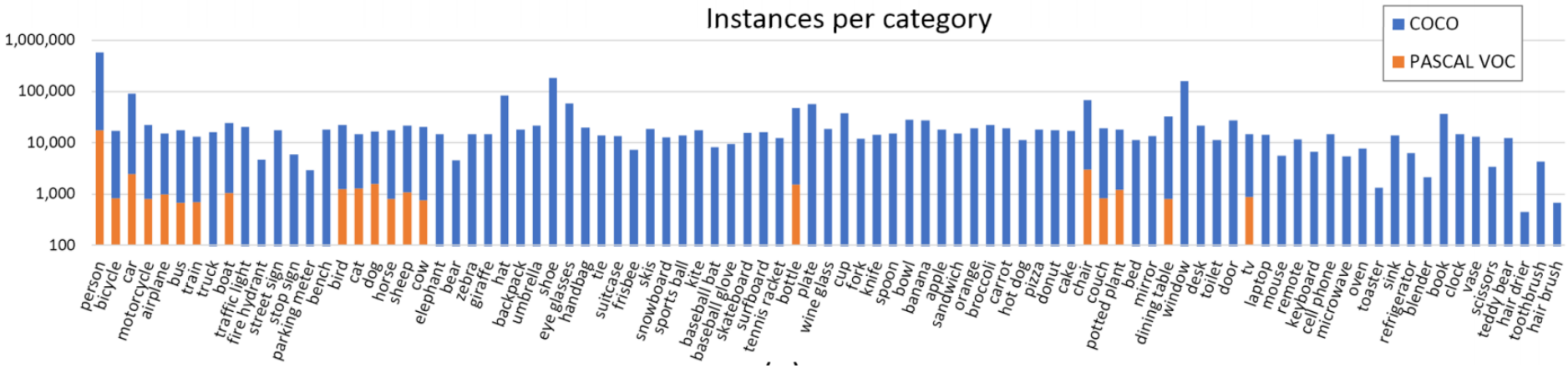
Microsoft CoCo[4]

- A baseball winds up to pitch the ball.
- A pitcher throwing the ball in a baseball game.
- A pitcher throwing a baseball on the mound.
- A baseball player pitching a ball on the mound.
- A left-handed pitcher throwing for the San Francisco giants.

- 120k train + validation images [vs 1k(Pascal), 31k(Flikr)].
- Instance level segmentations labels with 91 object classes and 2.5M labelled instances.
- Standard benchmark for image caption generation task.

[4] Lin et. al., Microsoft COCO: Common Objects in Context, 2014.[Dataset Link]

Microsoft CoCo[4]



Source: *Dataset Paper*

- 120k train + validation images [vs 1k(Pascal), 31k(Flikr)].
- Instance level segmentations labels with 91 object classes and 2.5M labelled instances.
- Standard benchmark for image caption generation task.

[4] Lin et. al., Microsoft COCO: Common Objects in Context, 2014.[Dataset Link]

Abstract Scenes Dataset[5]

Jenny loves to play soccer but she is worried that Mike will kick the ball too hard.

Mike and Jenny play outside in the sandbox. Mike is afraid of an owl that is in the tree.

Source: *L. Zitnick*

- 1002 sets of scenes with 10 images in each.
- Reduced variability (hence complexity) than real word scenes.
- Descriptions have non-visual attributes.
- Clip-arts provide segmentation labels.

[5] Zitnick et.al., Bringing Semantics Into Focus Using Visual Abstraction, 2013. [Dataset Link]

Abstract Scenes Dataset[5]

Source: *L. Zitnick*

- 1002 sets of scenes with 10 images in each.
- Reduced variability (hence complexity) than real word scenes.
- Descriptions have non-visual attributes.
- Clip-arts provide segmentation labels.

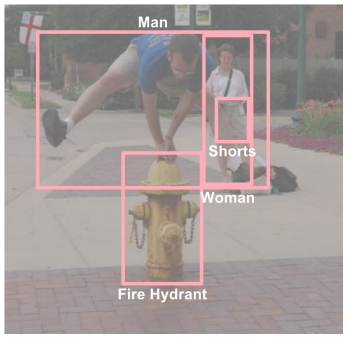
[5] Zitnick et.al., Bringing Semantics Into Focus Using Visual Abstraction, 2013. [Dataset Link]

Overview

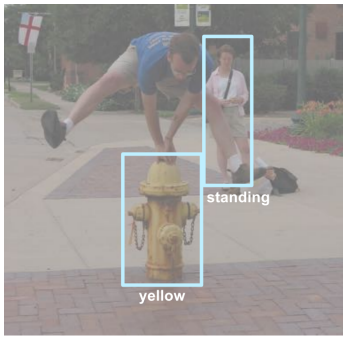
- Datasets for image caption generation
  - Single sentence generation
  - Multiple sentence/paragraph generation
- Datasets for video caption generation
- Datasets for referring expressions task
- Metrics
  - Image measures
  - Text measures
    - Automatic measures
    - Human based measures

Visual Genome Dataset[6]

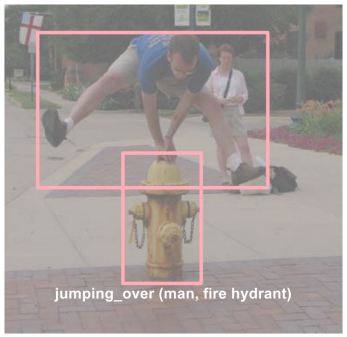
Objects



Attributes



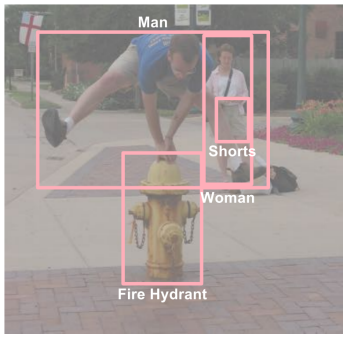
Relationships



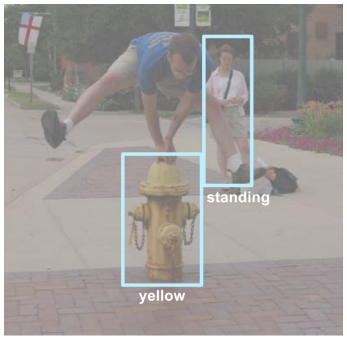
[6]Krishna et. al., *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*, 2016. [Dataset Link]

Visual Genome Dataset[6]

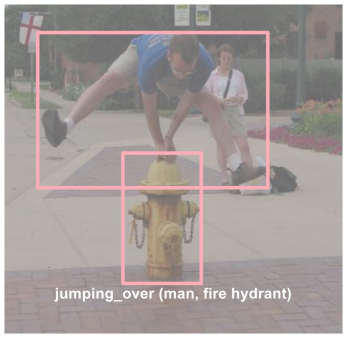
Objects

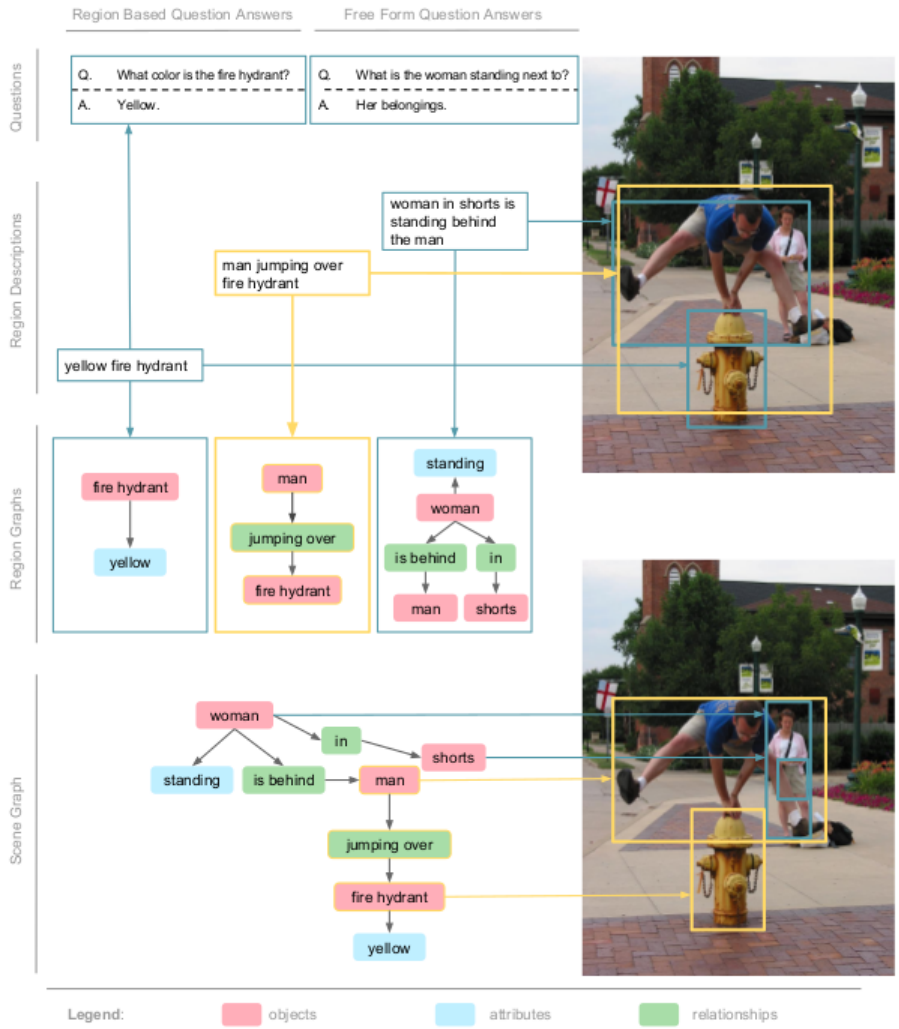


Attributes



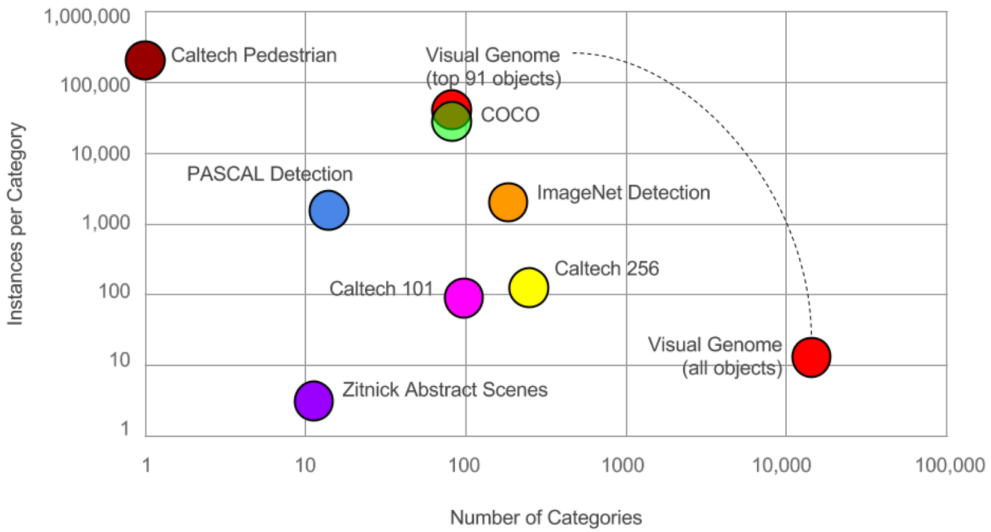
Relationships

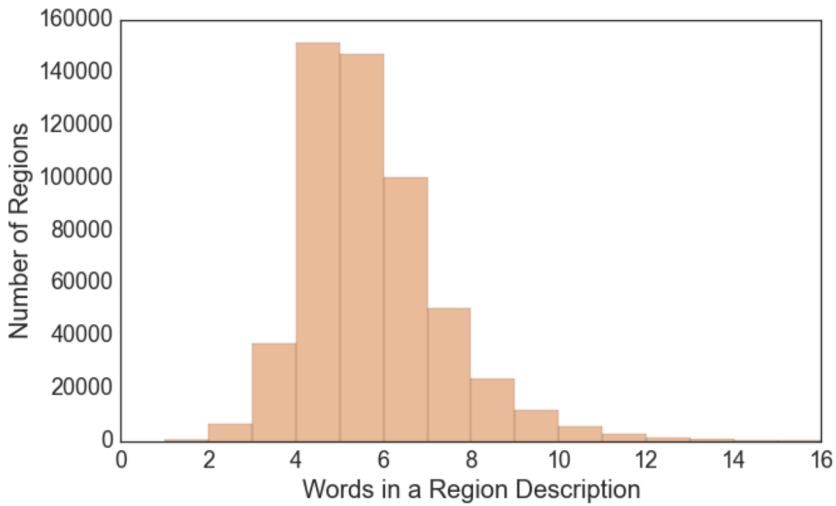




[6]Krishna et. al., *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*, 2016. [Dataset Link]

Visual Genome Dataset[6]





Source: *Dataset Paper*

Num. images	*Num. categories*	*Region desc./image*	*Objs./image*	*Attr./image*	*Rel./image*
∼ 108k	*∼ 18k*	*∼ 42*	*∼ 21*	*∼ 16*	*∼ 18*

Min. desc. length	*Max. desc. length*	*Word count/desc.*	*Objs./region*	*Attr./region*	*Rel./region*
1	*16*	*∼ 5*	*∼ 0.43*	*∼ 0.41*	*∼ 0.45*

[6]Krishna et. al., *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*, 2016. [Dataset Link]

Krause et al[7]

Sentences

1) A girl is eating donuts with a boy in a restaurant
2) A boy and girl sitting at a table with doughnuts.
3) Two kids sitting a coffee shop eating some frosted donuts
4) Two children sitting at a table eating donuts.
5) Two children eat doughnuts at a restaurant table.

Paragraph

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

Source: *Dataset paper*

[7] Krause et.al., A Hierarchical Approach for Generating Descriptive Image Paragraphs, 2016.

Krause et al[7]

Sentences

1) A girl is eating donuts with a boy in a restaurant
2) A boy and girl sitting at a table with doughnuts.
3) Two kids sitting a coffee shop eating some frosted donuts
4) Two children sitting at a table eating donuts.
5) Two children eat doughnuts at a restaurant table.

Paragraph

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

Source: *Dataset paper*

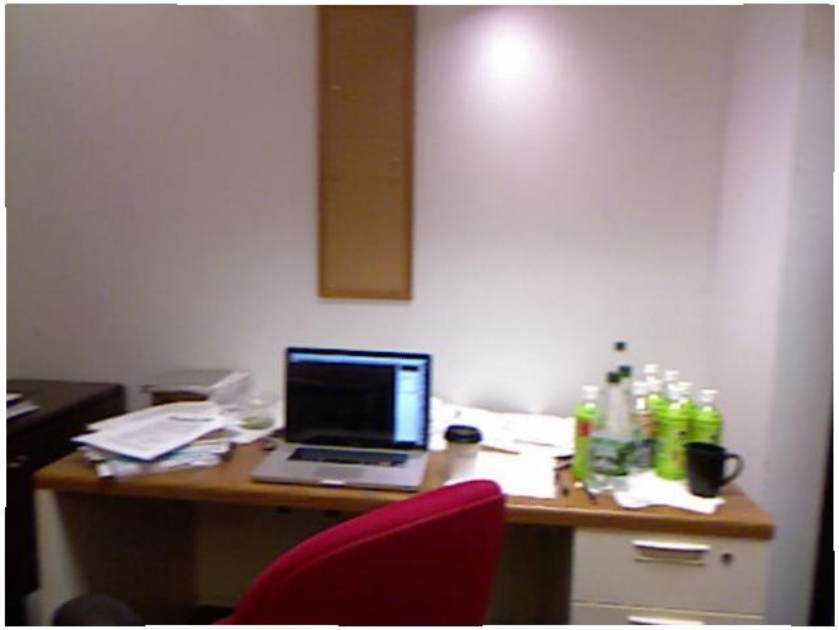
- *$\sim$ 20k images with following statistics (dataset to be public soon)*

Dataset	*Desc. Length*	*Sentence Length*	*Diversity**	*Nouns*	*Adj.*	*Verbs*	*Pro-nouns*
MS COCO	*11.30*	*11.30*	*19.01*	*33.45*	*27.23*	*10.72*	*1.23*
Krause et al	*67.50*	*11.91*	*70.49*	*25.81*	*27.64*	*15.21*	*2.45*

** Diversity = 100 - Avg. CIDER similarity among sentences for each image*

[7] Krause et.al., A Hierarchical Approach for Generating Descriptive Image Paragraphs, 2016.

Kong et al[8]



Description: A big office desk is in the middle of the room. A Mac laptop is on top of the desk. There are a few bottles on top of the desk, on the right of the laptop. In front of the bottles there is a blue mug.

Source: *S Fidler*

[8]Kong et.al., What are you talking about? Text-to-Image Coreference, 2014. [Dataset Link]

Kong et al[8]

Description: This room is filled with different types of furniture and home goods. The lights on the ceiling are strung across the room, they are circular and bright. At the back of the room, there are shelves filled with an assortment of pillows and blankets. There are a few couches facing away from those shelves. The couches have many pillows on top of them. On the second couch, which is dark green, sits a man in a plaid shirt. Another black couch faces the second couch. In front of the black couch is a shelf containing large brown bowls on the bottom shelf, towels on the second shelf, and vases on the top shelf. In front of the shelf is a dining table with brown wooden chairs, pink placemats, white dinnerware, and a brown glass bottle.

Source: *S Fidler*

[8]Kong et.al., What are you talking about? Text-to-Image Coreference, 2014. [Dataset Link]

Kong et al[8]

# sent	# words	min # sent	max sent	min words	max words
3.2	39.1	1	10	6	144

# nouns of interest	# pronouns	# scene mentioned	scene correct
3.4	0.53	0.48	83%

Table: Statistics **per description**.

- 1449 RGB-D images with 20 object categories.
- Long and complex descriptions.
- Significant co-reference.
- Deceiving information (object and scene mis-classification).

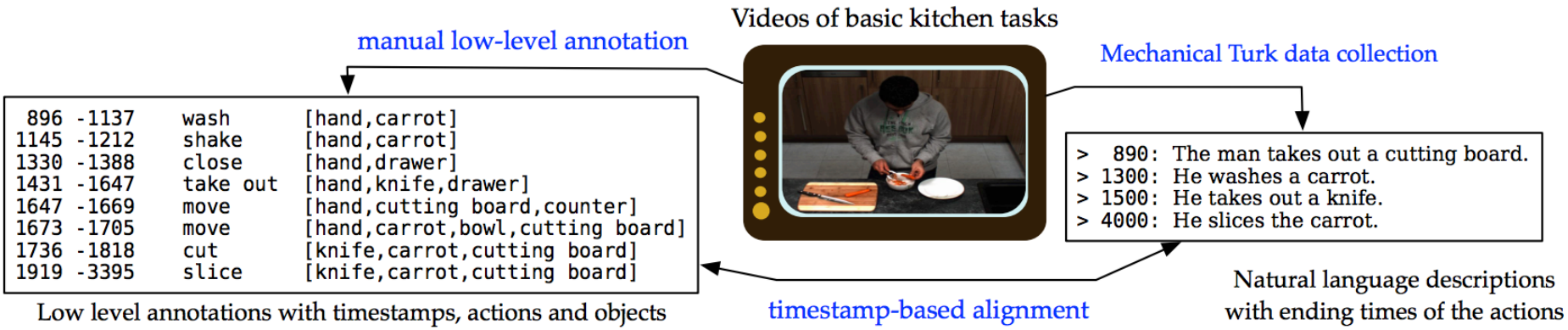
Source: *S Fidler*

[8]Kong et.al., What are you talking about? Text-to-Image Coreference, 2014. [Dataset Link]

Overview

- Datasets for image caption generation
  - Single sentence generation
  - Multiple sentence/paragraph generation
- Datasets for video caption generation
- Datasets for referring expressions task
- Metrics
  - Image measures
  - Text measures
    - Automatic measures
    - Human based measures

TACoS[9]



Source: *Michaela Regneri*

- 127 cooking videos with 20 different text descriptions/video.
- Time stamp labeling of textual descriptions with each description describing an activity label like `wash`, `slicing`, `trash`.
- Time stamp labelings of low level activity and participants(involving tool, patient, source, and target).
- Similarity scores of object activity pairs are available.

[9] Regneri et. al., Grounding Action Descriptions in Videos, 2013. [Dataset Link]

YouCook[10]

She chops the egg with an egg chopper and put the egg chopper in a glass container. Then she takes the egg mixture in the steel bowl and the bread pieces and butter which are kept in plates on the kitchen counter top. Then she places it near the sink. Then she applies butter on the frying pan and takes the chopped egg kept in the steel bowl.

- 88 videos with ∼8 descriptions/video.
- Each video annotated with human descriptions, tracks for 48 different objects (belonging to 7 categories), and time intervals of 7 different actions.

[10] Das et. al., A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching, 2013. [Dataset Link]

LSMDC[11]

He lights a match.

Spinning wearily in his chair, the director faces SOMEONE.

A missile explodes nearby. A few blocks away, the boys run past houses and across more lawns.

On the street, a helmeted soldier talks into a transceiver.

Source: *Anna Rohrbach*

- Audio descriptions and script data aligned with movie videos.

[11] Rohrbach et. al., Movie Description, 2017. [Dataset Link]

LSMDC[11]

- Audio descriptions (ADs) are descriptions for the visually impaired, prepared by trained describers and professional narrators.
- Usually ADs have better visual descriptions and accurate alignment than script data.
- Movies have more diversity and realistic than cooking videos.
- Statistics:

Num. movies	Num. clips	Num. sentences	Avg. length of clip	Avg. sentences / clip
200	128k	128k	4.1s	> 1

[11] Rohrbach et. al., Movie Description, 2017. [Dataset Link]

LSMDC[12]

Tasks

- Movie description
  - Generate a single sentence to describe a given clip.
- Movie annotation and retrieval
  - There are two tracks (Multiple Choice Test and Movie Retrieval)
  - In the Multiple Choice test, a video clip with 5 possible captions is given. And the correct caption needs to be determined.
  - In the Retrieval task, given a text query, the nearest video needs to be retrieved.
  - MovieQA[11] dataset has a similar Multiple Choice task, but more information (movie clips, plots, subtitles and ADs) can be used to determine the correct caption.
- Fill in the blanks
  - Given a clip and a sentence with a blank, the task is to fill that blank.

---

[11] Tapaswi et. al., MovieQA: Understanding Stories in Movies through Question-Answering, 2016. [Dataset Link]

[12] Rohrbach et. al., Movie Description, 2017. [Dataset Link]

Overview

- Datasets for image caption generation
  - Single sentence generation
  - Multiple sentence/paragraph generation
- Datasets for video caption generation
- Datasets for referring expressions task
- Metrics
  - Image measures
  - Text measures
    - Automatic measures
    - Human based measures

Referring Expressions Dataset

RefClef	RefCOCO	RefCOCO+
		
right rocks rocks along the right side stone right side of stairs	woman on right in white shirt woman on right right woman	guy in yellow dirbbling ball yellow shirt and black shorts yellow shirt in focus

- This task involves referring to the particular objects described in natural language.
- Based on the ReferIt task, with more descriptive language expressions.
- Several datasets[13, 14] have been concurrently developed for this task.

[13] Yu et. al., Modeling Context in Referring Expressions, 2016. [Dataset Link]

[14] Mao et. al., Generation and Comprehension of Unambiguous Object Descriptions, 2016. [Dataset Link]

Overview

- Datasets for image caption generation
  - Single sentence generation
  - Multiple sentence/paragraph generation
- Datasets for video caption generation
- Datasets for referring expressions task
- Metrics
  - Image measures
  - Text measures
    - Automatic measures
    - Human based measures

Image Measures

- IoU[15](or Jaccard Index)

$$\text{IoU}\,(A, B) = \frac{A \cap B}{A \cup B}$$

- Precision, Recall, F1 measure

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

[15] Everingham et al

BLEU[17] (BiLingual Evaluation Understudy)

a : candidate sentence, b : set of reference sentences, w_n : n-gram

$c_x(y_n)$: count of n-gram y_n in sentence x.

- Based on n-gram based precision.
- $\text{BLEU}_n(a, b) = \dfrac{\sum_{w_n \in a} \min\left(c_a(w_n), \max_{j=1,\ldots,|b|} c_{b_j}(w_n)\right)}{\sum_{w_n \in a} c_a(w_n)}$
- BLEU or $\text{BLEU}_{\text{Overall}}$ is a geometric mean of n-gram scores from 1 to 4.

[16] Detailed results in: Callison-Burch et. al., 2006; Reiter et. al., 2008; Hodosh et. al., 2013

[17] Papineni et. al., BLEU: A Method for Automatic Evaluation of Machine Translation, 2002

BLEU[17] (BiLingual Evaluation Understudy)

a : candidate sentence, b : set of reference sentences, w_n : n-gram

$c_x(y_n)$: count of n-gram y_n in sentence x.

- Based on n-gram based precision.
- $$\texttt{BLEU}_n(a,b) = \frac{\sum_{w_n \in a} \min\left(c_a(w_n), \max_{j=1,\ldots,|b|} c_{b_j}(w_n)\right)}{\sum_{w_n \in a} c_a(w_n)}$$
- BLEU or $\texttt{BLEU}_{\text{Overall}}$ is a geometric mean of n-gram scores from 1 to 4.
- Strength
  - Automatic, easy to compute
- Weakness[16]
  - No constraints on the ordering of n-grams.
  - Each n-gram is treated equally.
  - A measure of fluency rather than semantic similarity between a and b.

[16] Detailed results in: Callison-Burch et. al., 2006; Reiter et. al., 2008; Hodosh et. al., 2013

[17] Papineni et. al., BLEU: A Method for Automatic Evaluation of Machine Translation, 2002

Rouge[18] (Recall Oriented Understudy of Gisting Evaluation)

a : candidate sentence, b : set of reference sentences, w_n : n-gram
$c_x(y_n)$: count of n-gram y_n in sentence x.

- Based on n-gram based recall.
- $\texttt{ROUGE}_n(a, b) = \dfrac{\sum_{j=1}^{|b|} \sum_{w_n \in b_j} \min\left(c_a(w_n), c_{b_j}(w_n)\right)}{\sum_{j=1}^{|b|} \sum_{w_n \in b_j} c_{b_j}(w_n)}$

[18] Lin et. al., ROUGE: A Package for Automatic Evaluation of Summaries, 2004

Rouge[18] (Recall Oriented Understudy of Gisting Evaluation)

a : candidate sentence, b : set of reference sentences, w_n : n-gram

$c_x(y_n)$: count of n-gram y_n in sentence x.

- Based on n-gram based recall.
- $$\texttt{ROUGE}_n(a, b) = \frac{\sum_{j=1}^{|b|} \sum_{w_n \in b_j} \min\left(c_a(w_n), c_{b_j}(w_n)\right)}{\sum_{j=1}^{|b|} \sum_{w_n \in b_j} c_{b_j}(w_n)}$$
- There are other variants like $\texttt{ROUGE}_S$, $\texttt{ROUGE}_L$

[18] Lin et. al., ROUGE: A Package for Automatic Evaluation of Summaries, 2004

Rouge[18] (Recall Oriented Understudy of Gisting Evaluation)

a : candidate sentence, b : set of reference sentences, w_n : n-gram

$c_x(y_n)$: count of n-gram y_n in sentence x.

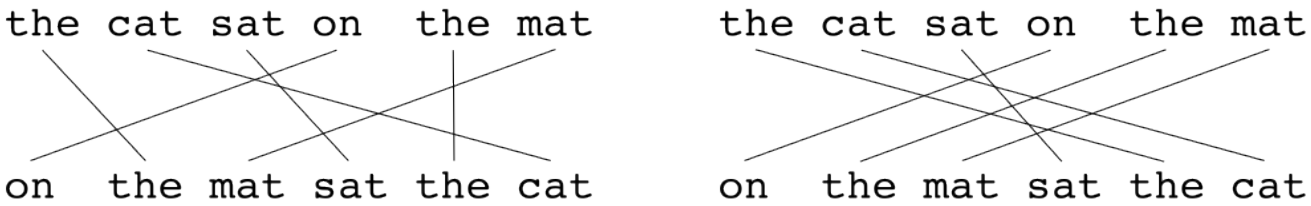
- Based on n-gram based recall.
- $$\texttt{ROUGE}_n(a, b) = \frac{\sum_{j=1}^{|b|} \sum_{w_n \in b_j} \min\left(c_a(w_n), c_{b_j}(w_n)\right)}{\sum_{j=1}^{|b|} \sum_{w_n \in b_j} c_{b_j}(w_n)}$$
- There are other variants like $\texttt{ROUGE}_S$, $\texttt{ROUGE}_L$
- Similar strengths and weaknesses as BLEU.

[18] Lin et. al., ROUGE: A Package for Automatic Evaluation of Summaries, 2004

METEOR[19] (Metric for Evaluation of Translation with Explicit ORdering)

a : candidate sentence, b : set of reference sentences

- An alignment between a and b is first computed.



\# criss-crosses of left alignment is less

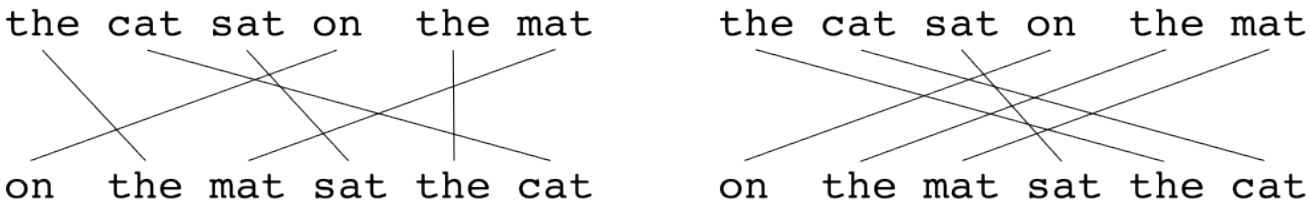
Source: *Wikipedia*

[19] Banerjee et. al., METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, 2005

METEOR[19] (Metric for Evaluation of Translation with Explicit ORdering)

a : candidate sentence, b : set of reference sentences

- An alignment between a and b is first computed.



Source: *Wikipedia*

criss-crosses of left alignment is less

- $\texttt{METEOR} = \max_{j=1,\ldots,|b|} \left(\frac{10PR}{R+9P}\right)\left(1-\frac{1}{2}\left(\frac{\#\text{chunks}}{\#\text{matched unigrams}}\right)^3\right)$

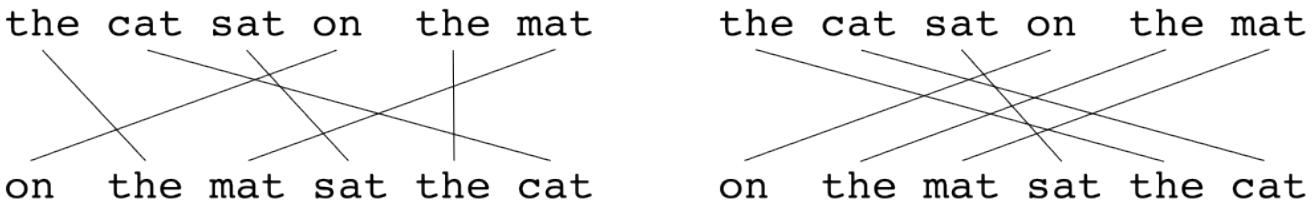
  P = unigram precision, R = unigram recall,
  chunks = set of unigrams adjacent in a and b_j (Example on the right has 3 chunks).

[19] Banerjee et. al., METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, 2005

METEOR[19] (Metric for Evaluation of Translation with Explicit ORdering)

a : candidate sentence, b : set of reference sentences

- An alignment between a and b is first computed.



Source: *Wikipedia*

\# criss-crosses of left alignment is less

- $\texttt{METEOR} = \max_{j=1,\ldots,|b|} \left(\frac{10PR}{R+9P}\right)\left(1-\frac{1}{2}\left(\frac{\#\text{chunks}}{\#\text{matched unigrams}}\right)^3\right)$

  P = unigram precision, R = unigram recall,
  chunks = set of unigrams adjacent in a and b_j (Example on the right has 3 chunks).
- Smoother penalization of different ordering of chunks.
- Higher correlation with human consensus scores.

[19] Banerjee et. al., METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, 2005

CIDEr[20]

a : candidate sentence, b : set of reference sentences

- $CIDEr_n(a, b) = \frac{1}{|b|} \sum_{j=1}^{|b|} \frac{\mathbf{g^n}(a) \cdot \mathbf{g^n}(b_j)}{\|\mathbf{g^n}(a)\| \|\mathbf{g^n}(b_j)\|}$

  $\mathbf{g^n}(x)$: vector formed by TF-IDF scores of all n-grams in x.

  $CIDEr(a, b) = \sum_{n=1}^{N} w_n CIDEr_n(a, b)$

[20] Vedantam et. al., CIDEr: Consensus-based Image Description Evaluation, 2014

CIDEr[20]

a : candidate sentence, b : set of reference sentences

- $$CIDEr_n(a,b) = \frac{1}{|b|}\sum_{j=1}^{|b|}\frac{\mathbf{g^n}(a)\cdot\mathbf{g^n}(b_j)}{\left\|\mathbf{g^n}(a)\right\|\left\|\mathbf{g^n}(b_j)\right\|}$$

  $\mathbf{g^n}(x)$: vector formed by TF-IDF scores of all n-grams in x.

  $$CIDEr(a,b) = \sum_{n=1}^{N} w_n CIDEr_n(a,b)$$
- Gives more weight-age to *important* n-grams.
- Higher correlation with human consensus scores compared to above metrics.

[20] Vedantam et. al., CIDEr: Consensus-based Image Description Evaluation, 2014

SPICE[21]

Motivation

‘False positive’
(High n-gram similarity)

A young girl *standing on top of a* tennis court.

A giraffe *standing on top of a* green field.

‘False negative’
(Low n-gram similarity)

A shiny metal pot filled with some diced veggies.

The pan on the stove has chopped vegetables in it.

...n-gram overlap is not necessary or sufficient for two sentences to mean the same

...SPICE primarily addresses false positives

Source: *Peter Anderson*

[21] Anderson et. al., SPICE: Semantic Propositional Image Caption Evaluation, 2016

SPICE[22]

"A young girl standing on top of a basketball court"

Semantic propositions:

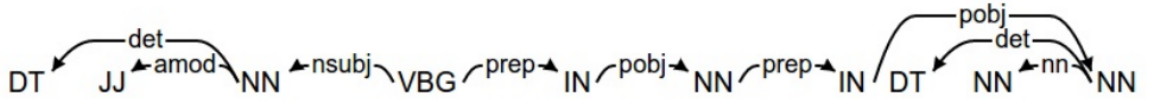
1. There is girl
2. The girl is young
3. The girl is standing
4. There is court
5. The court is used for basketball
6. The girl is on the court

Source: *Peter Anderson*

[22] Anderson et. al., SPICE: Semantic Propositional Image Caption Evaluation, 2016

Key Idea – scene graphs[1]

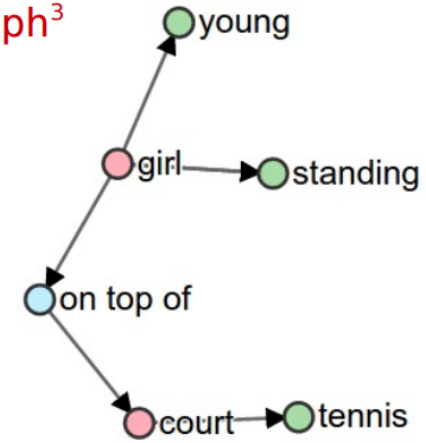
2. Parse[2]



1. Input

A young girl standing on top of a tennis court

3. Scene Graph[3]



4. Tuples

(girl)
(court)
(girl, young)
(girl, standing)
(court, tennis)
(girl, on-top-of, court)

[1] Johnson et. al. Image Retrieval Using Scene Graphs, CVPR 2015
[2] Klein & Manning: Accurate Unlexicalized Parsing, ACL 2003
[3] Schuster et. al: Generating semantically precise scene graphs from textual descriptions for improved image retrieval, EMNLP 2015

Source: *Peter Anderson*

[23] Anderson et. al., SPICE: Semantic Propositional Image Caption Evaluation, 2016

SPICE[24]

SPICE calculated as an F-score over tuples, with:

- Merging of synonymous nodes, and
- Wordnet synsets used for tuple matching and merging.

Given candidate caption *c*, a set of reference captions *S*, and the mapping *T* from captions to tuples:

$$P(c, S) = \frac{|T(c) \otimes T(S)|}{|T(c)|}$$

$$R(c, S) = \frac{|T(c) \otimes T(S)|}{|T(S)|}$$

$$SPICE(c, S) = F_1(c, S) = \frac{2 \cdot P(c, S) \cdot R(c, S)}{P(c, S) + R(c, S)}$$

Source: *Peter Anderson*

[24] Anderson et. al., SPICE: Semantic Propositional Image Caption Evaluation, 2016

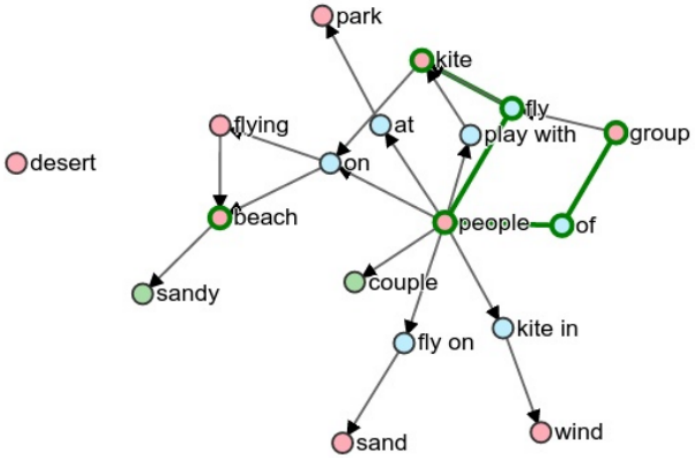
SPICE[25]

Example

Reference captions

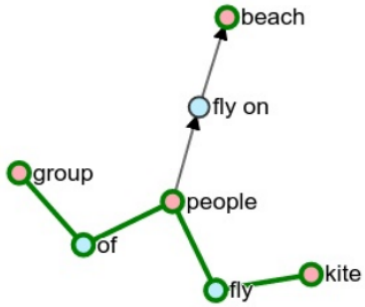
"People playing with kites outside in the desert."
"A group of people at a park flying a kite. "
"A group of people flying a kite on a sandy beach"
"People on the beach flying kites in the wind."
"A couple people out flying a kite on some sand."

Reference scene graph



Candidate caption & scene graph

"a group of people flying kites on a beach"



SPICE F-Score: 0.429, Pr: 0.857, Re: 0.286

Source: *Peter Anderson*

[25] Anderson et. al., SPICE: Semantic Propositional Image Caption Evaluation, 2016

SPICE[26]

- Pros:
  - Places importance on capturing details about objects, attributes and relationships.
  - Higher correlation with humans compared to n-gram based metrics.

- Cons:
  - This metric does not check whether the grammar is correct.
  - Depends on semantic parsers, which might not always be correct.
  - Equal weighting of different nouns, attributes, relationships.

[26] Anderson et. al., SPICE: Semantic Propositional Image Caption Evaluation, 2016

Ranking based measures

- **Recall@k** = % image sentence pairs for which the ground truth sentence was present in the top-k list.
- **Median rank** = k at which the system has a recall of 50%.

[27] Hodosh et. al., Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, 2013

Ranking based measures

- **Recall@k** = % image sentence pairs for which the ground truth sentence was present in the top-k list.
- **Median rank** = k at which the system has a recall of 50%.
- Such measures can be used for retrieval based systems.

[27] Hodosh et. al., Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, 2013

Ranking based measures

- **Recall@k** = % image sentence pairs for which the ground truth sentence was present in the top-k list.
- **Median rank** = k at which the system has a recall of 50%.
- Such measures can be used for retrieval based systems.
- Hodosh et. al.[27] shows that both automatic ranking based measures are *more* robust than metrics that consider only the quality of the first result.

[27] Hodosh et. al., Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, 2013

Human based measures

- Measuring quality of a single best result
  - **Rating system of 1-4** from Hodosh et. al.[28]

[28] Hodosh et. al., Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, 2013

[29] Manning et. al., Introduction to Information Retrieval, 2008

Human based measures

- Measuring quality of a single best result
  - **Rating system of 1-4** from Hodosh et. al.[28]
- Measuring ranked candidates
  - **Success@k** = % image sentence pairs for which at least one relevant result is found in the top-k list.
  - **R-precision**[29] = average % of relevant items in the top-k list.

[28] Hodosh et. al., Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics, 2013

[29] Manning et. al., Introduction to Information Retrieval, 2008

Challenges involving humans

- Datasets
  - Using humans to make binary/choosing decisions, rather than complex decisions. Helps in faster and quality annotation.[30]
  - Games to make the creation of datasets more interesting for annotators.[31]
  - Anyhow involves post-processing to remove spelling mistakes, and sometimes grammatical mistakes.

[30] Parikh et. al., 2011; Vedantam et. al., 2014

[31] Deng et. al., 2013; Kazemzadeh et. al., 2014

[32] More details in Reiter et. al., 2008; Hodosh et. al., 2013

Challenges involving humans

- Datasets
  - Using humans to make binary/choosing decisions, rather than complex decisions. Helps in faster and quality annotation.[30]
  - Games to make the creation of datasets more interesting for annotators.[31]
  - Anyhow involves post-processing to remove spelling mistakes, and sometimes grammatical mistakes.
- Metrics
  - Hodosh et. al. used qualification tests to get *experts* to compare correlation between human based measures and automatic measures.[32]
  - Common practice to use averaged responses from humans rather than single responses. Vedantam et. al.(2014) uses as many as 50 human responses per image sentence pair to ensure the quality of responses.

[30] Parikh et. al., 2011; Vedantam et. al., 2014

[31] Deng et. al., 2013; Kazemzadeh et. al., 2014

[32] More details in Reiter et. al., 2008; Hodosh et. al., 2013