Knowledge-aware image and video captioning

Spencer Whitehead

srw5@illinois.edu

Acknowledgement: Portions of this content borrowed from Abigail See, Chris Manning, and Tamara Berg

Outline

- Overview of NLG
    - Summarization
- Image and video captioning
- Knowledge-aware Captioning
- Evaluating NLG

Outline

- **Overview of NLG**
  - Summarization
- Image and video captioning
- Knowledge-aware Captioning
- Evaluating NLG

Overview of NLG

- NLG ≈ generating new text
- NLG is a component of
  - Dialogue
  - (Free-form) QA
  - (Abstractive) Summarization
  - Machine Translation
  - Captioning
  - …

Overview of NLG

- Language modeling
  - Predict next word given previous words

    $P(y_t | y_1,\ldots,y_{t-1})$

- Language model = system that produces this probability distribution
  - RNN
  - Tranformer
  - …

Overview of NLG

- Conditional language modeling
  - Predict next word given previous words and some other input, x.

    $P(y_t | y_1,\ldots,y_{t-1},x)$

- Conditional language modeling tasks:
  - Machine translation
  - Summarization
  - Dialogue

Summarization

- Given input text, x, generate summary, y, such that y is shorter than x and y contains main information of x.

- Single-document:
  - x is one document
    - News article

- Multi-document:
  - Input is set of documents $\{x_1,...,x_n\}$
  - Documents typically topically related

Summarization

Extractive

- Select parts of original text to form summary (no generation)
- Rigid
- Easier in some settings

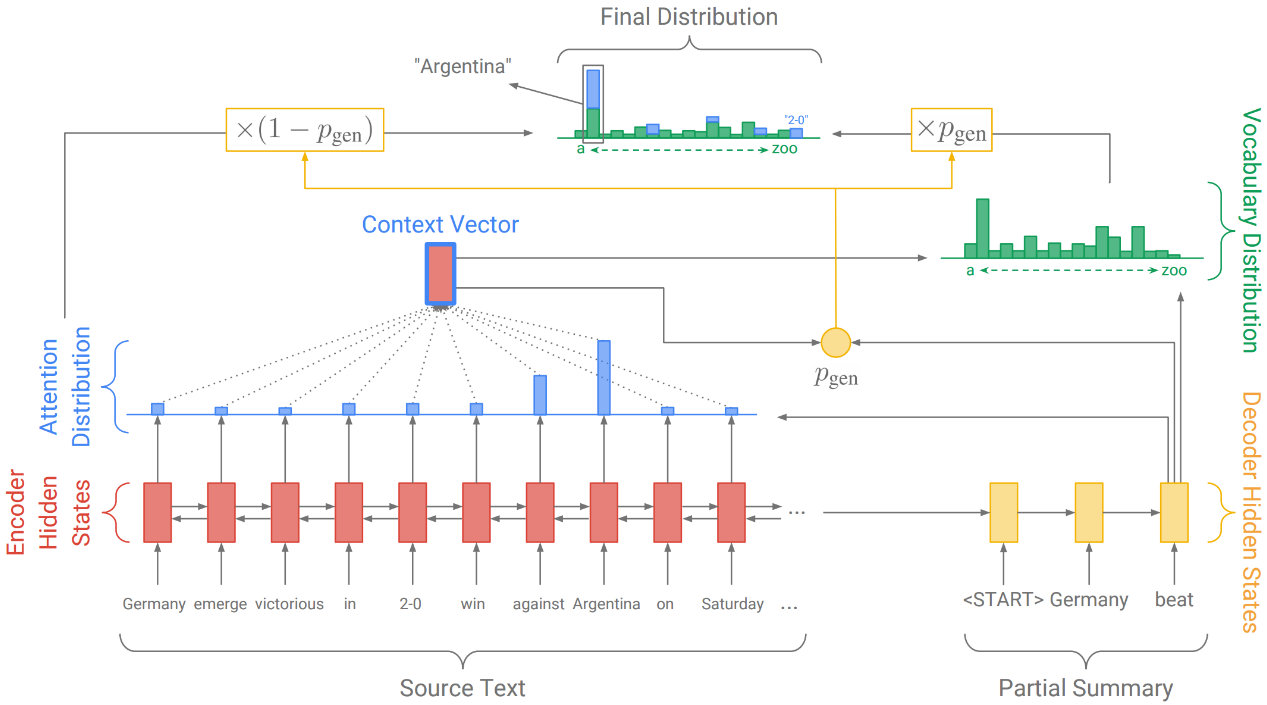
Abstractive

- Generate (mostly) new summary text using NLG techniques
- Flexible
- More difficult

Summarization

- Copy mechanism
  - Seq2seq+attention models
    - Good at writing fluent output
    - Bad a getting details correct
  - Copy mechanisms
    - Enable seq2seq system to copy words/phrases from input
    - Neural models that can generate and copy are very useful
      - Copy details, generate the rest
- Gu et al., 2016: https://arxiv.org/pdf/1603.06393.pdf
  - Copy mechanism in MT for rare words
- Nallapatti et al., 2016: https://arxiv.org/pdf/1602.06023.pdf
  - Copy mechanism in abstractive summarization

Summarization (One Example)



- On each decoder step, calculate p_{gen} (probability of generating a word from the vocabulary, rather than copying it). Final distribution is mixture of generation and copy (attention) probabilities:

$$P(w) = p_{gen}P_{vocab}(w) + (1 – p_{gen})\ \Sigma_{z:z=w}a_z^t$$

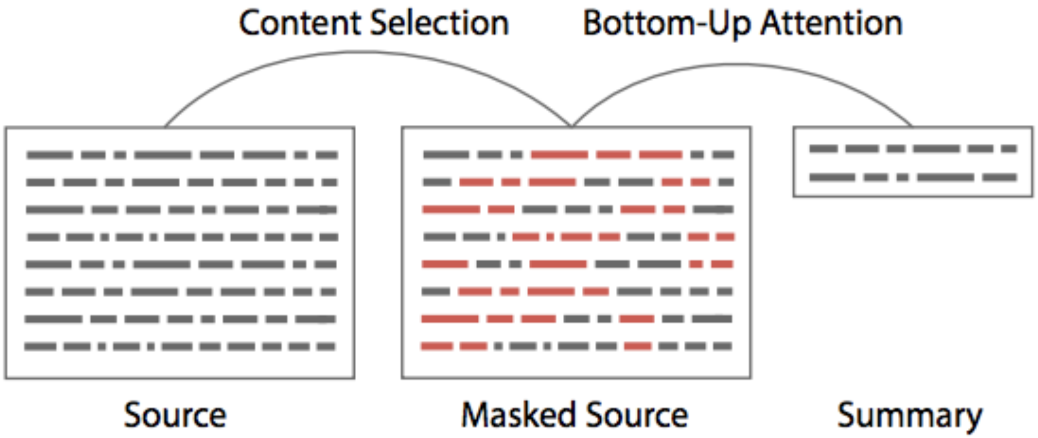
See et al., 2017: https://arxiv.org/pdf/1704.04368.pdf

Summarization

- Problems with copy mechanisms
  - Copy too much
    - Abstractive but almost extractive
  - Bad content selection
    - No specific strategy for selecting what to copy
- Pre-neural techniques separated content selection and realization
  - Seq2seq mixes these together → No content selection strategy

Summarization

- Solution: Bottom-up Summarization
  - Gehrmann et al., 2018: https://arxiv.org/pdf/1808.10792.pdf
- Content selection stage:
  - Use sequence tagging model to select which words to include or not include
- Bottom-up attention stage:
  - Mask words that should not be included
- Simple yet effective
  - Less copying



Outline

- Overview of NLG
    - Summarization
- **Image and video captioning**
- Knowledge-aware Captioning
- Evaluating NLG

Image Captioning

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2

More Nuanced than Recognition

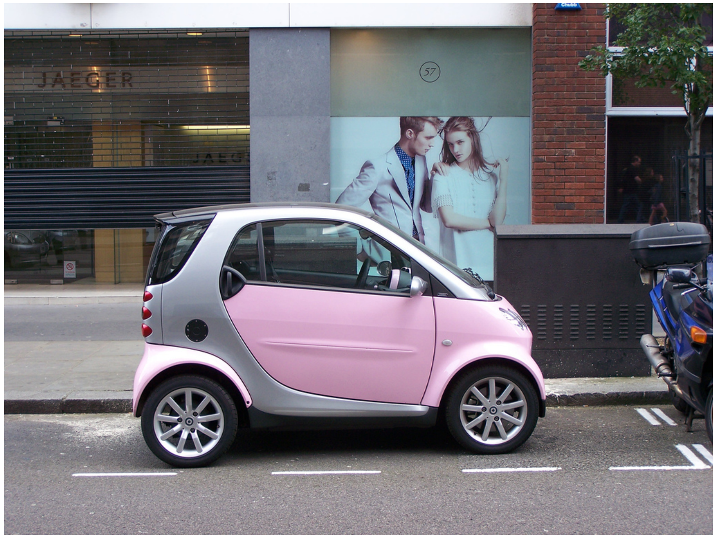
person

car

shoe

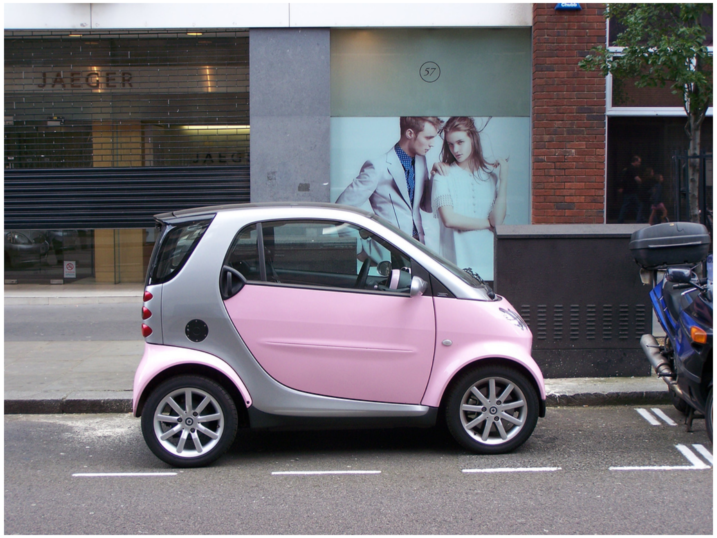
Berg, Attributes Tutorial CVPR13

Towards Complex Structured Outputs



car

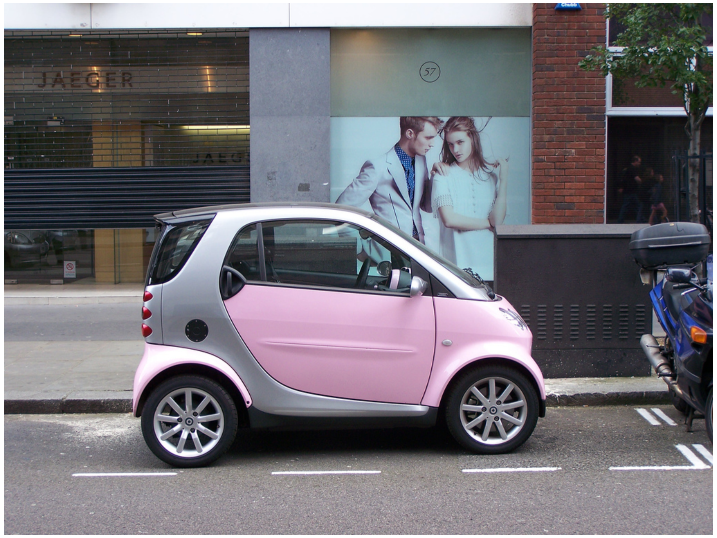
Towards Complex Structured Outputs



pink car

Attributes of objects

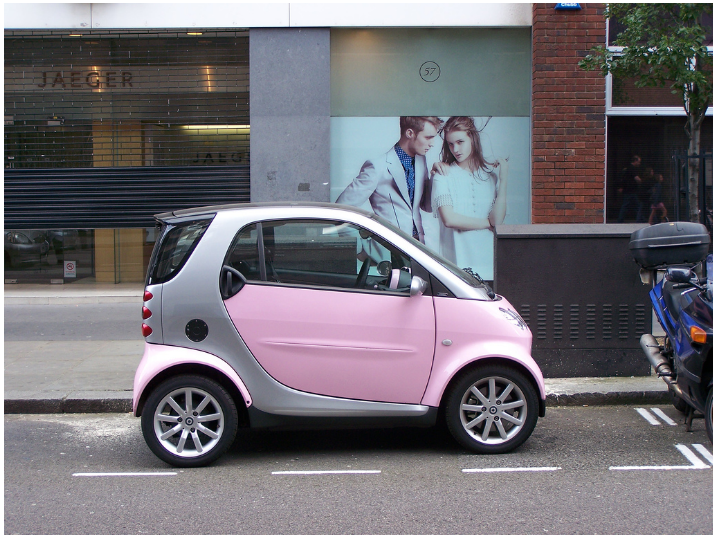
Towards Complex Structured Outputs



pink car on the road

Relationships between objects

Towards Complex Structured Outputs



Little pink smart car parked on the side of a road in a London shopping district.

… Complex structured recognition outputs

Telling the *"story of an image"*

Learning from Descriptive Text

"It was an arresting face, pointed of chin, square of jaw. Her eyes were pale green without a touch of hazel, starred with bristly black lashes and slightly tilted at the ends. Above them, her thick black brows slanted upward, cutting a startling oblique line in her magnolia-white skin–that skin so prized by Southern women and so carefully guarded with bonnets, veils and mittens against hot Georgia suns"

Scarlett O'Hara described in Gone with the Wind.

How does the world work?

Visually descriptive language provides:

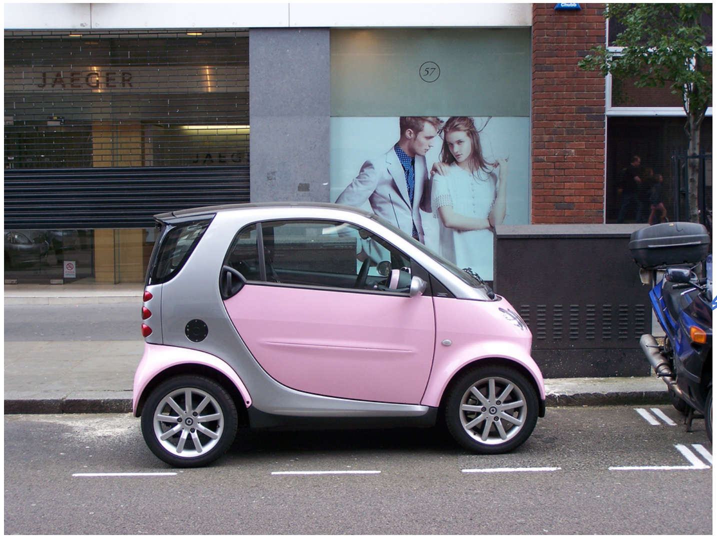
- Information about the world, especially the visual world.
- information about how people construct natural language for imagery.
- guidance for visual recognition.

What should we recognize?

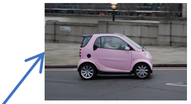
How do people describe the world?

Berg, Attributes Tutorial CVPR13

Methodology



Pink Car
Sign
Door
Motorcycle
Tree
Brick building
Dirty Road
Sidewalk
London
Shopping district



A random Pink Smart Car seen driving around Lambeth Roundabout and onto Lambeth Bridge.

Smart Car. It was so adorable and cute in the parking lot of the post office, I had to stop and take a picture.

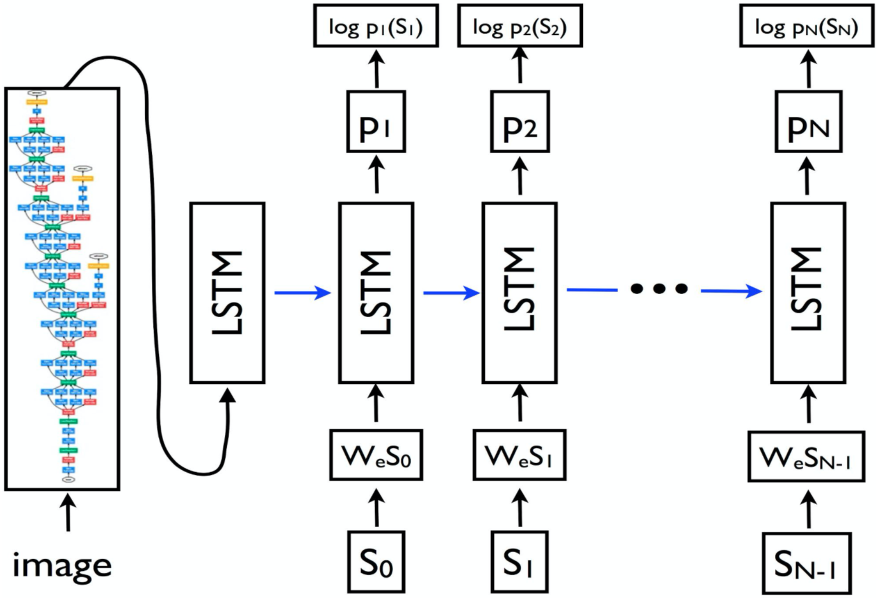
Natural language description

Generation Methods:

1) Compose descriptions directly from recognized content
2) Retrieve relevant existing text given recognized content

Image Captioning

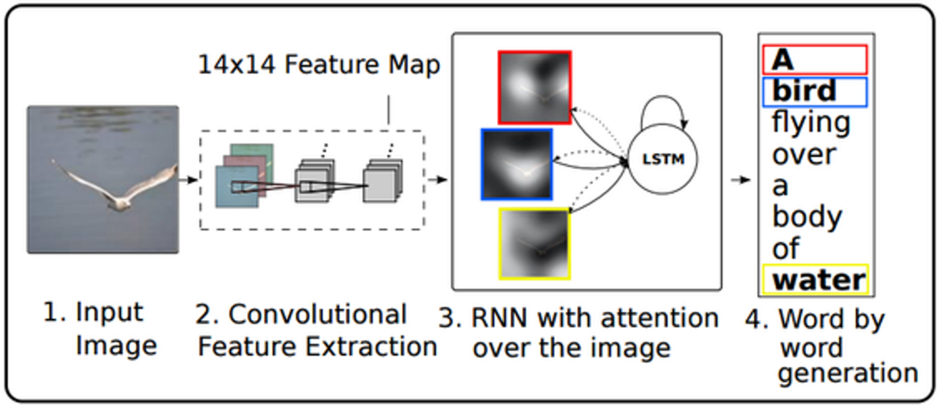
- Follow encoder-decoder model
  - CNN encoder
- Input image into CNN
- Use response from fully connected layer as initial state



Vinyals et al., 2015: https://arxiv.org/abs/1411.4555

Image Captioning

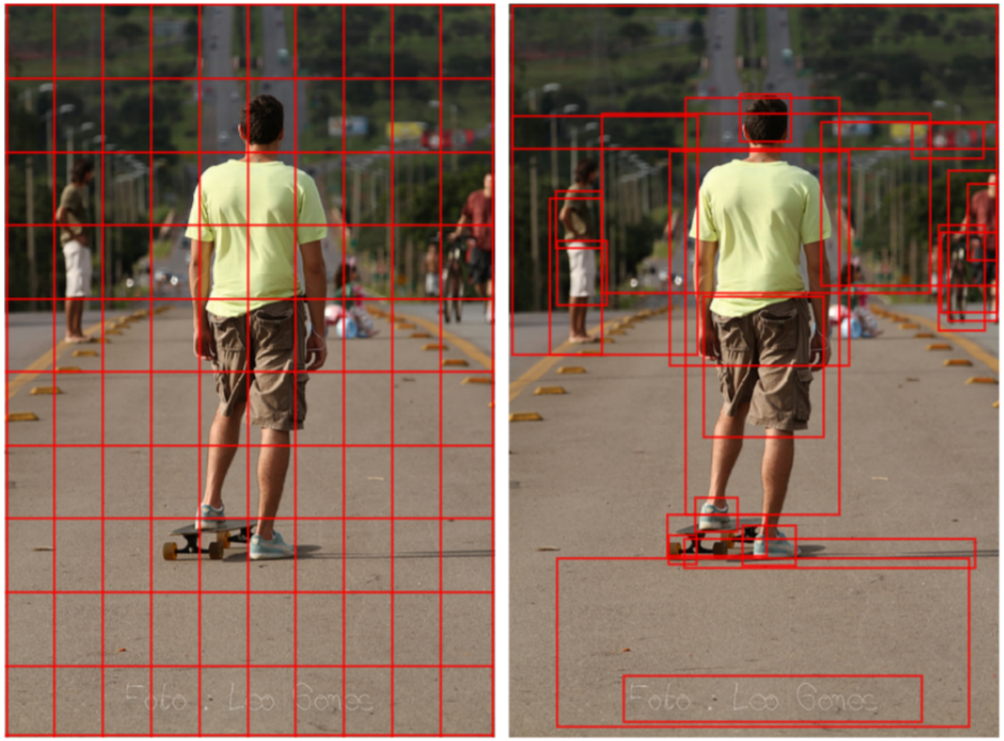
- Visual attention (Xu et al., 2015)
- Again, encoder-decoder
  - CNN encoder
  - RNN decoder
- Use convolutional feature map
- Attend on different locations


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Xu et al., 2015: https://arxiv.org/abs/1502.03044

Image Captioning

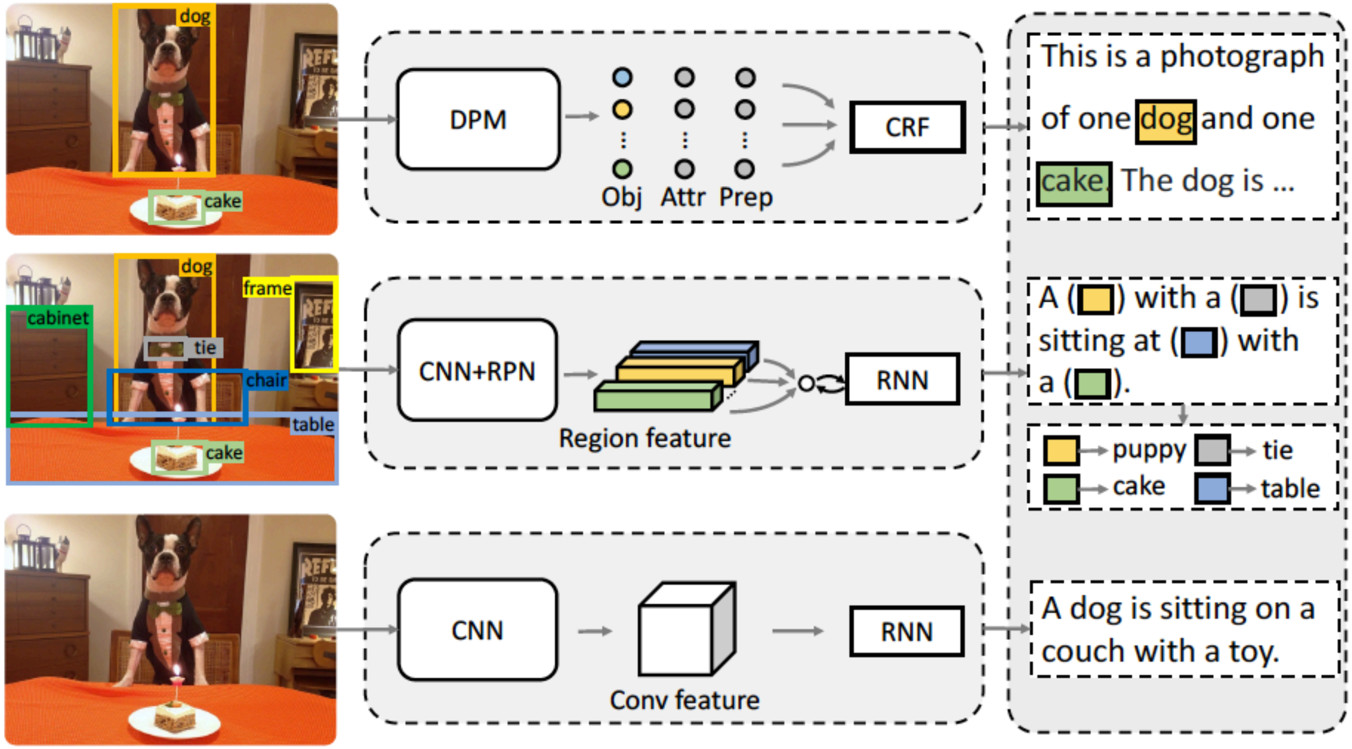
- Bottom-Up and Top-Down Attention (Anderson et al.,2018)
    - Attend over objects + salient regions



Anderson et al., 2018: https://arxiv.org/abs/1707.07998

Image Captioning

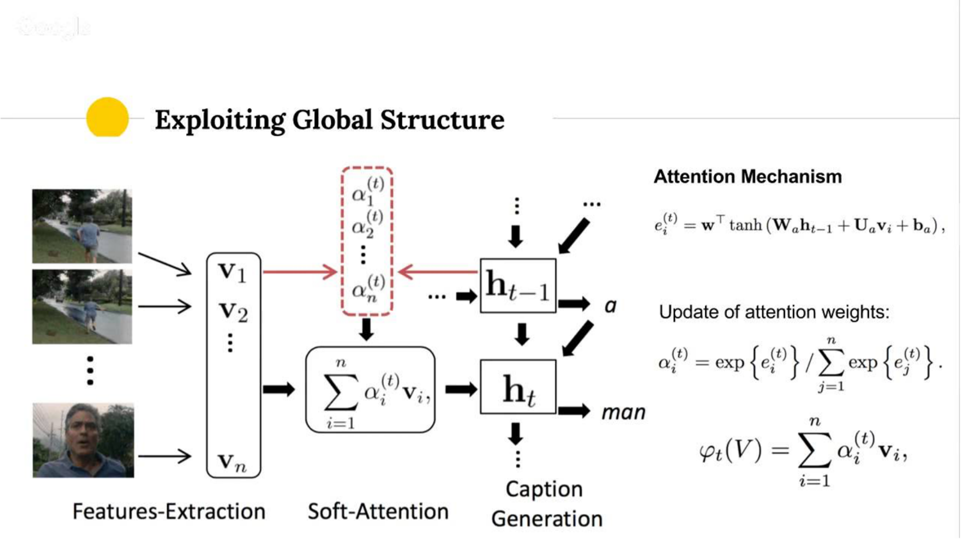
- Neural Baby Talk (Lu et al., 2018)
  - Generate template caption, point to objects


dog
cake
DPM
Obj
Attr
Prep
CRF
This is a photograph of one dog and one cake. The dog is …
dog
frame
cabinet
tie
chair
table
cake
CNN+RPN
Region feature
RNN
A (🟨) with a (⬜) is sitting at (🟦) with a (🟩).
🟨 → puppy
⬜ → tie
🟩 → cake
🟦 → table
CNN
Conv feature
RNN
A dog is sitting on a couch with a toy.

Lu et al., 2018: https://arxiv.org/abs/1803.09845

Video Captioning

- Input video frames into encoder and decode (Venugopalan et al., 2015a: https://arxiv.org/abs/1505.00487)
- Average frame representations (Venugopalan et al., 2015b: https://arxiv.org/abs/1412.4729)
- Temporal attention (Yao et al., 2015: https://arxiv.org/abs/1502.08029)



Exploiting Global Structure

Attention Mechanism

$$e_i^{(t)} = \mathbf{w}^\top \tanh\left(\mathbf{W}_a \mathbf{h}_{t-1} + \mathbf{U}_a \mathbf{v}_i + \mathbf{b}_a\right),$$

Update of attention weights:

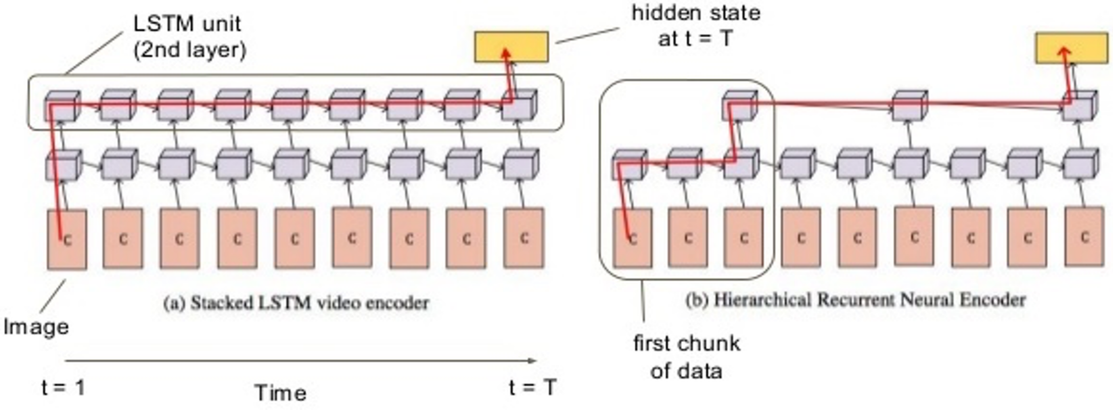
$$\alpha_i^{(t)} = \exp\left\{e_i^{(t)}\right\} / \sum_{j=1}^{n} \exp\left\{e_j^{(t)}\right\}.$$

$$\varphi_t(V) = \sum_{i=1}^{n} \alpha_i^{(t)} \mathbf{v}_i,$$

Yao et al., 2015: https://arxiv.org/abs/1502.08029

Video Captioning

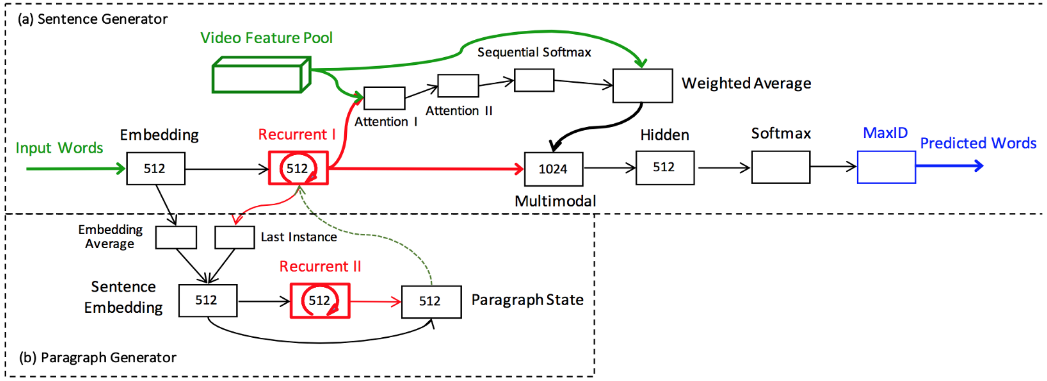
- Hierarchical approaches



(a) Stacked LSTM video encoder

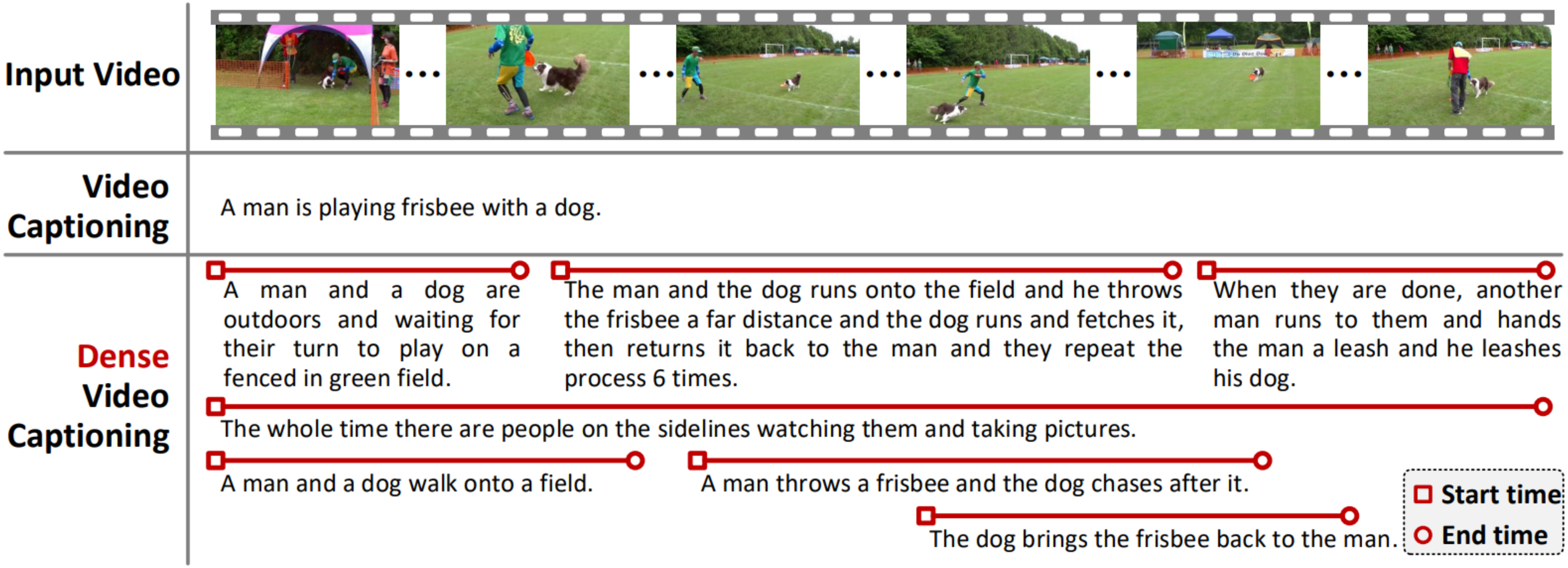
(b) Hierarchical Recurrent Neural Encoder

Pan et al., 2016:https://arxiv.org/abs/1511.03476



Yu et al., 2016: https://arxiv.org/abs/1510.07712

Video Captioning



A man is playing frisbee with a dog.

A man and a dog are outdoors and waiting for their turn to play on a fenced in green field.

The man and the dog runs onto the field and he throws the frisbee a far distance and the dog runs and fetches it, then returns it back to the man and they repeat the process 6 times.

When they are done, another man runs to them and hands the man a leash and he leashes his dog.

The whole time there are people on the sidelines watching them and taking pictures.

A man and a dog walk onto a field.

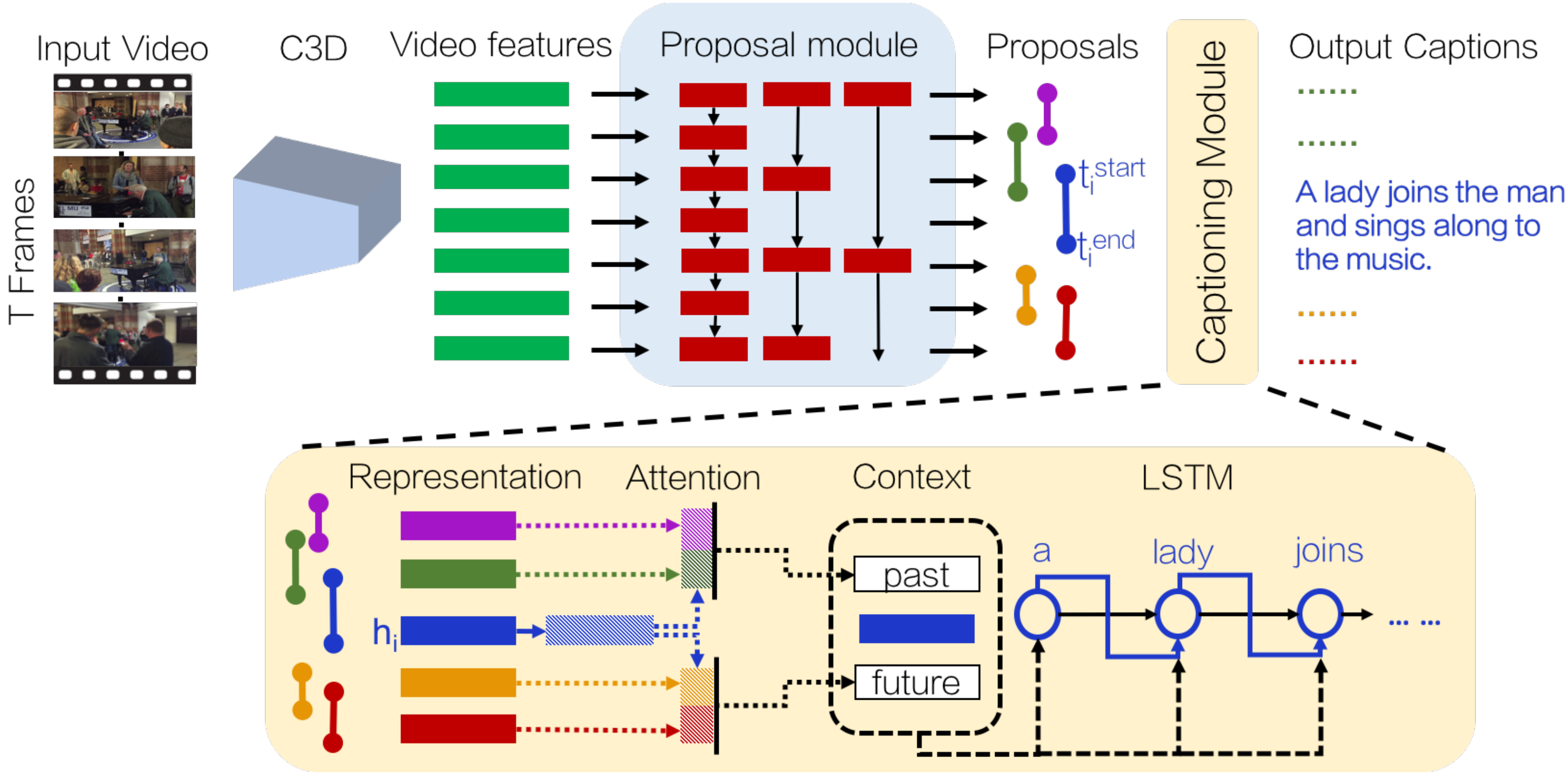
A man throws a frisbee and the dog chases after it.

The dog brings the frisbee back to the man.

Li et al., 2018: https://arxiv.org/abs/1804.08274

Video Captioning

- Dense video captioning (Krishna et al., 2017)



Krishna et al., 2017: https://arxiv.org/abs/1705.00754

Outline

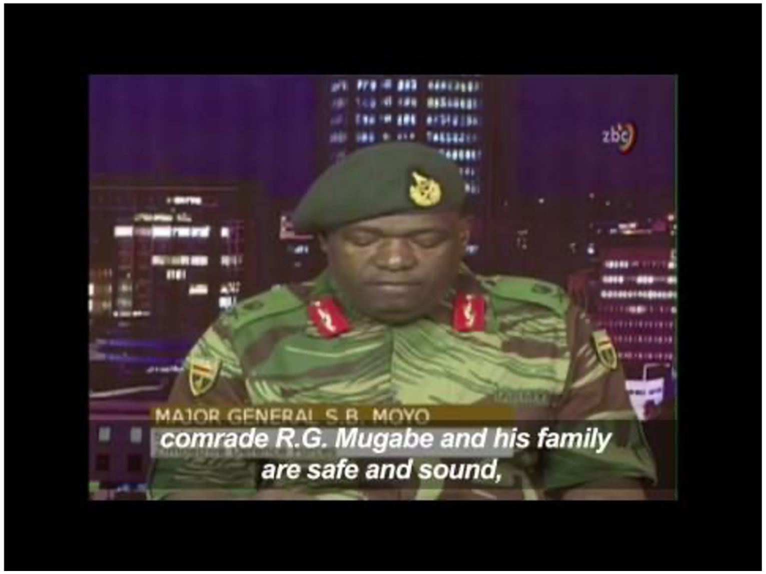
- Overview of NLG
    - Summarization
- Image and video captioning
- **Knowledge-aware Captioning**
- Evaluating NLG

What is Knowledge-aware NLG?

Knowledge-aware Captioning

- Given an image/video, generate a natural language description (typically a single sentence), which describes the ***context behind and contents of*** the video.

Ex: Spokesperson for Zimbabwe Defence Forces denies military coup.



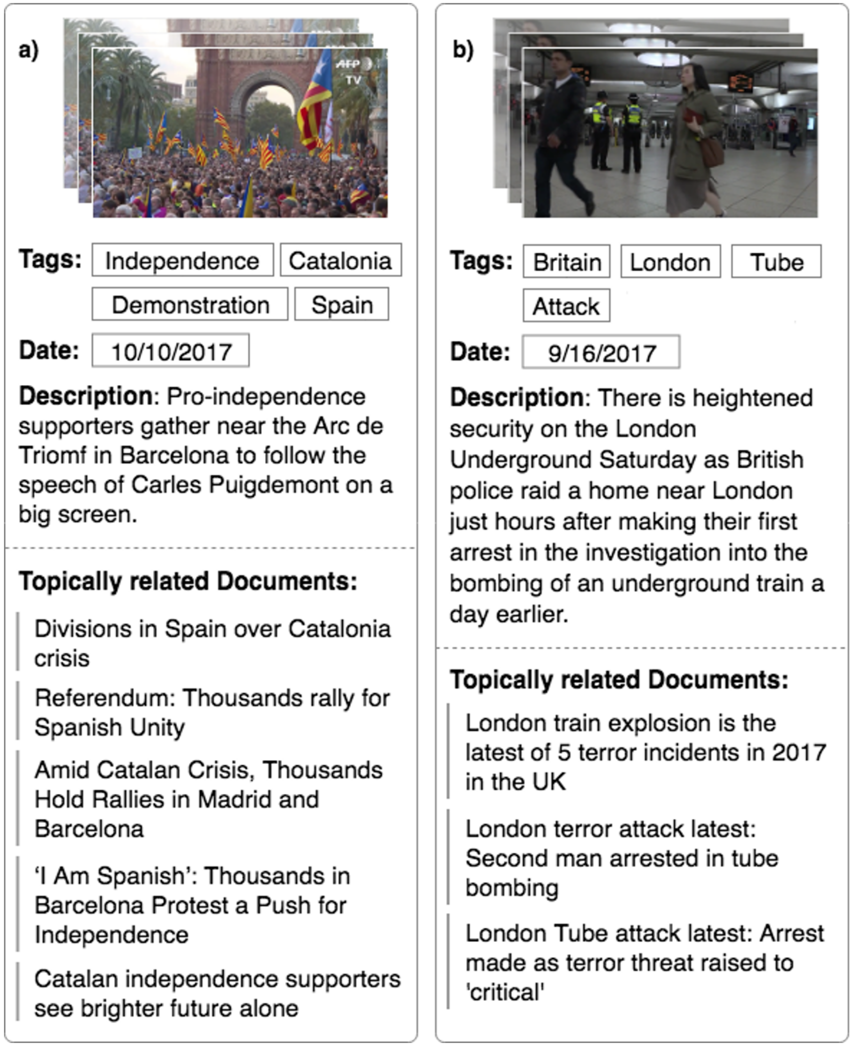
AFP News

Ex: A man is talking.

Chen and Dolan, 2011: https://www.aclweb.org/anthology/P11-1020/

Knowledge-aware (Video) Captioning

- How do we generate specific entity names and events?
- Visual evidence alone is insufficient
  - Named entities → low probabilities
  - Large scale visual recognition
    - Hard to train
    - Limited to famous entities
- If videos come with metadata, retrieve background documents
- Background documents alone is insufficient



a)

Tags: Independence | Catalonia | Demonstration | Spain

Date: 10/10/2017

Description: Pro-independence supporters gather near the Arc de Triomf in Barcelona to follow the speech of Carles Puigdemont on a big screen.

Topically related Documents:

Divisions in Spain over Catalonia crisis

Referendum: Thousands rally for Spanish Unity

Amid Catalan Crisis, Thousands Hold Rallies in Madrid and Barcelona

'I Am Spanish': Thousands in Barcelona Protest a Push for Independence

Catalan independence supporters see brighter future alone

b)

Tags: Britain | London | Tube | Attack

Date: 9/16/2017

Description: There is heightened security on the London Underground Saturday as British police raid a home near London just hours after making their first arrest in the investigation into the bombing of an underground train a day earlier.

Topically related Documents:

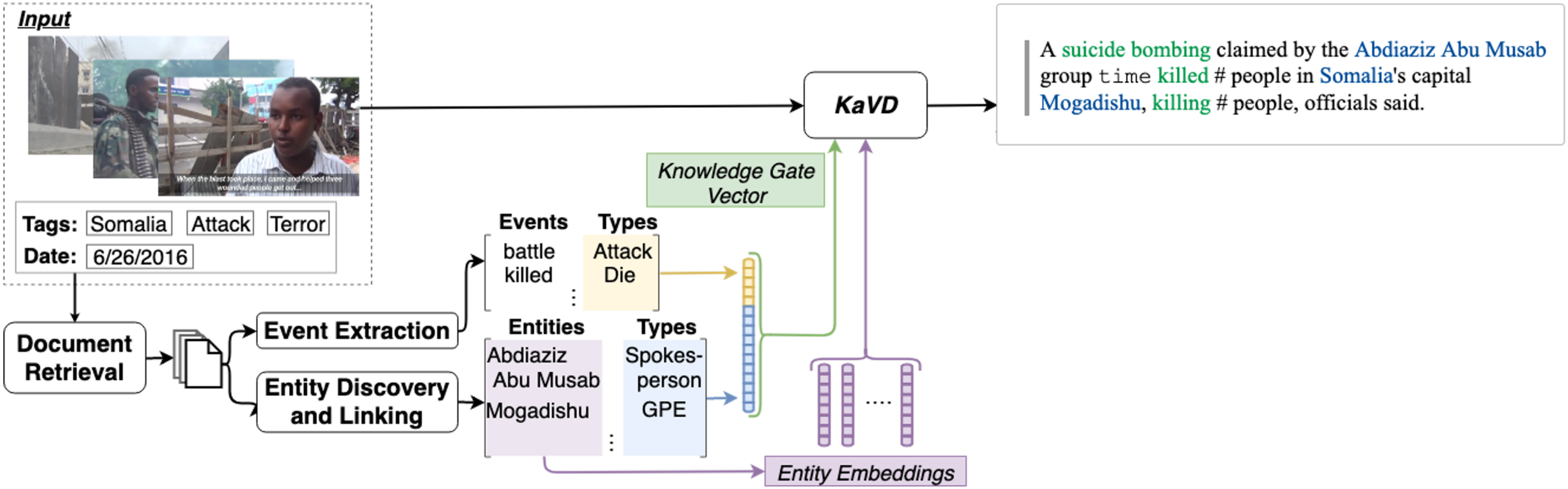
London train explosion is the latest of 5 terror incidents in 2017 in the UK

London terror attack latest: Second man arrested in tube bombing

London Tube attack latest: Arrest made as terror threat raised to 'critical'

Whitehead et al., 2018: https://www.aclweb.org/anthology/D18-1433/

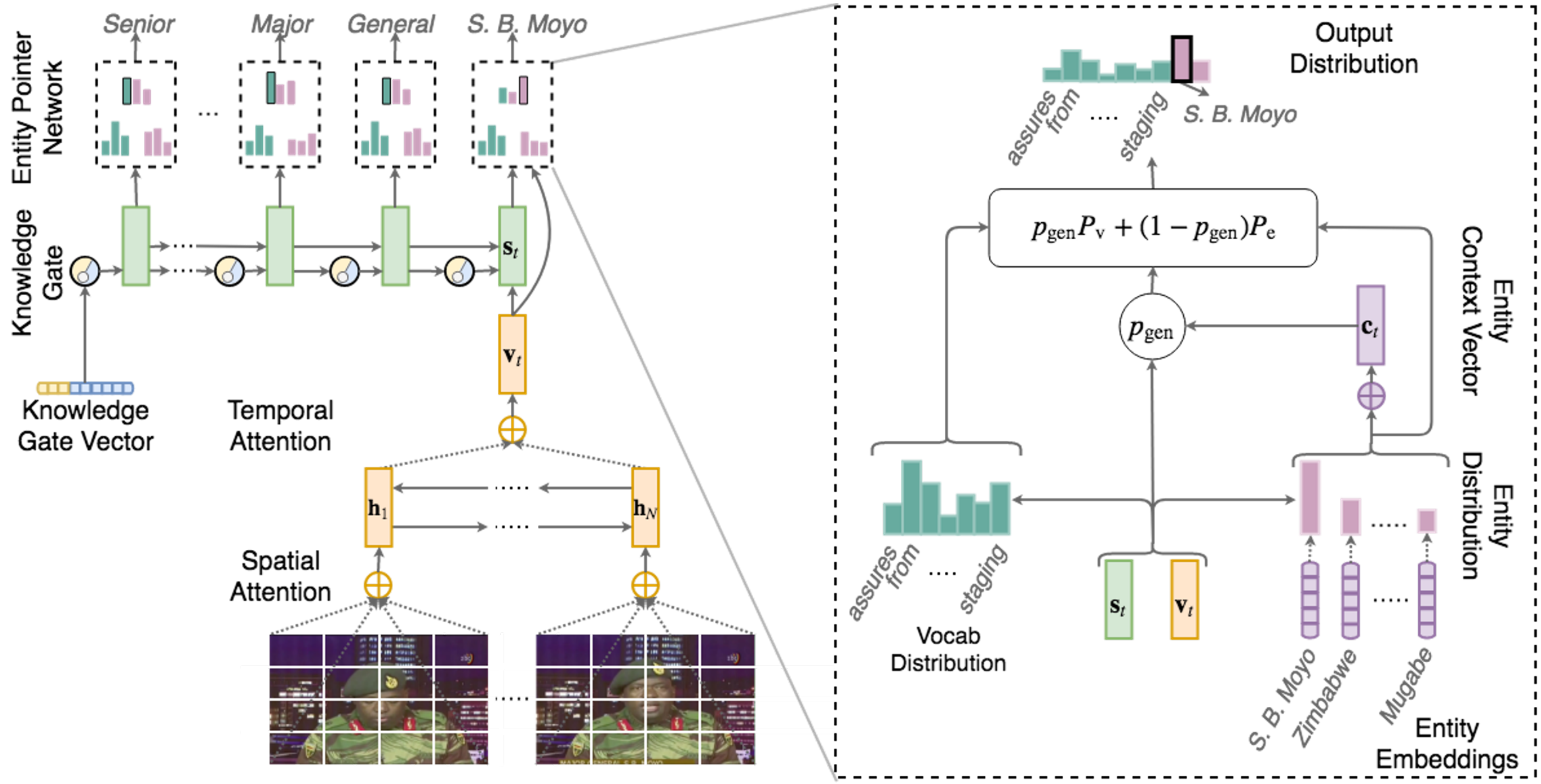
Knowledge-aware (Video) Captioning



A suicide bombing claimed by the Abdiaziz Abu Musab group `time` killed # people in Somalia's capital Mogadishu, killing # people, officials said.

Whitehead et al., 2018: https://www.aclweb.org/anthology/D18-1433/

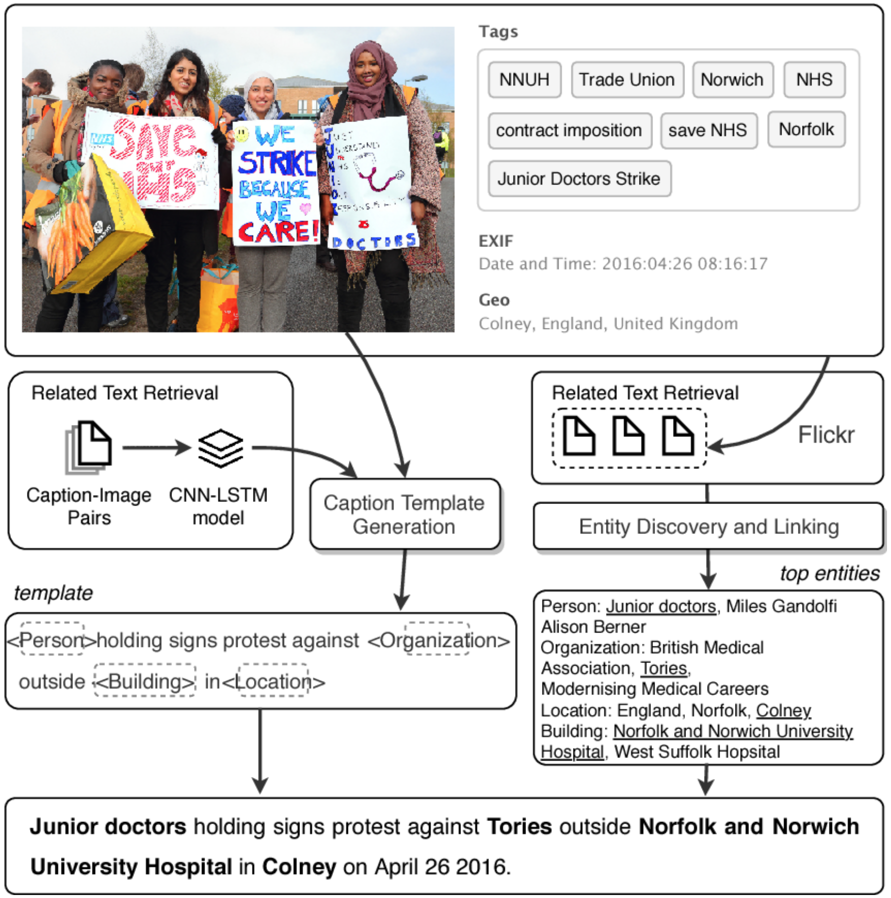
Knowledge-aware (Video) Captioning



Whitehead et al., 2018: https://www.aclweb.org/anthology/D18-1433/

Knowledge-aware (Image) Captioning

- Dataset
  - 34k news images for training and validation
    - Flickr
  - 2.5k social media images for evaluation
    - Reuters
- Condensed caption to ease generation
  - Limited semantic drift



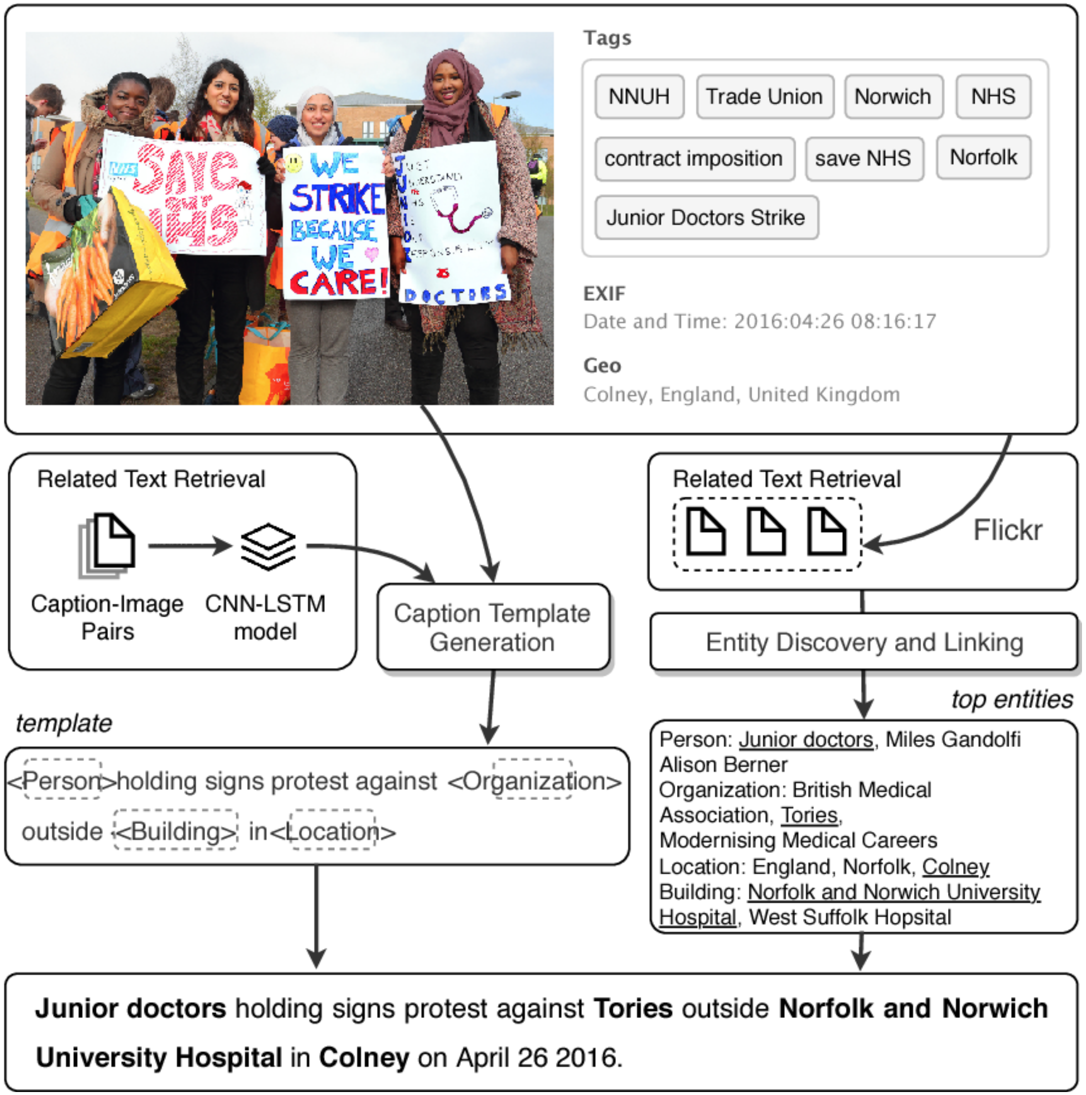
<Person> holding signs protest against <Organization> outside <Building> in <Location>

Person: Junior doctors, Miles Gandolfi Alison Berner
Organization: British Medical Association, Tories, Modernising Medical Careers
Location: England, Norfolk, Colney
Building: Norfolk and Norwich University Hospital, West Suffolk Hopsital

Junior doctors holding signs protest against **Tories** outside **Norfolk and Norwich University Hospital** in **Colney** on April 26 2016.

Lu et al., 2018: https://www.aclweb.org/anthology/D18-1435/

Knowledge-aware (Image) Captioning



Tags

NNUH | Trade Union | Norwich | NHS | contract imposition | save NHS | Norfolk | Junior Doctors Strike

EXIF

Date and Time: 2016:04:26 08:16:17

Geo

Colney, England, United Kingdom

Related Text Retrieval

Caption-Image Pairs → CNN-LSTM model

Caption Template Generation

Related Text Retrieval

Flickr

Entity Discovery and Linking

template

<Person> holding signs protest against <Organization> outside <Building> in <Location>

top entities

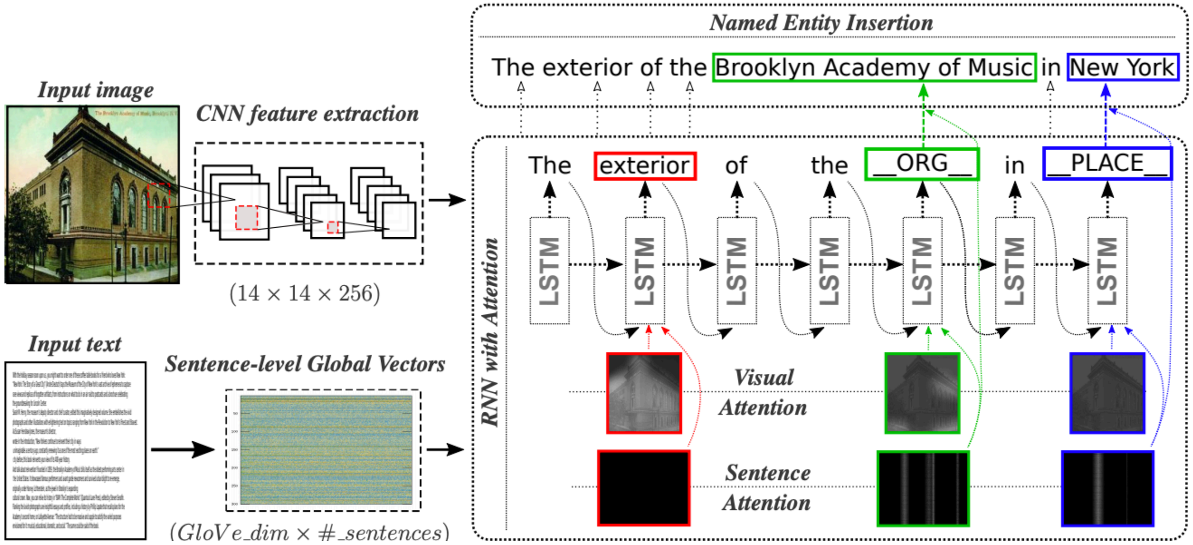
Person: Junior doctors, Miles Gandolfi Alison Berner
Organization: British Medical Association, Tories, Modernising Medical Careers
Location: England, Norfolk, Colney
Building: Norfolk and Norwich University Hospital, West Suffolk Hopsital

Junior doctors holding signs protest against **Tories** outside **Norfolk and Norwich University Hospital** in **Colney** on April 26 2016.

Lu et al., 2018: https://www.aclweb.org/anthology/D18-1435/

Knowledge-aware (Image) Captioning

- GoodNews dataset (Biten et al., 2019: https://arxiv.org/abs/1904.01475)
    - Similar to Lu et al. (2018) → template-based generation
    - Encodes entire article
    - Bigger dataset: 466k



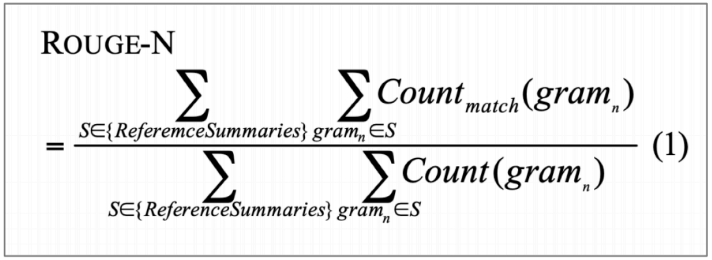
Outline

- Overview of NLG
  - Summarization
- Image and video captioning
- Knowledge-aware Captioning
- **Evaluating NLG**

Evaluating NLG

- BLEU
  - Papineni et al., 2002: https://www.aclweb.org/anthology/P02-1040
  - Precision-based, n-gram overlap
    - Important for MT

- ROUGE
  - Lin, 2004: http://www.aclweb.org/anthology/W04-1013
  - Recall based, n-gram overlap
    - Important for summarization



$$\text{ROUGE-N} = \frac{\sum_{S\in\{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S\in\{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

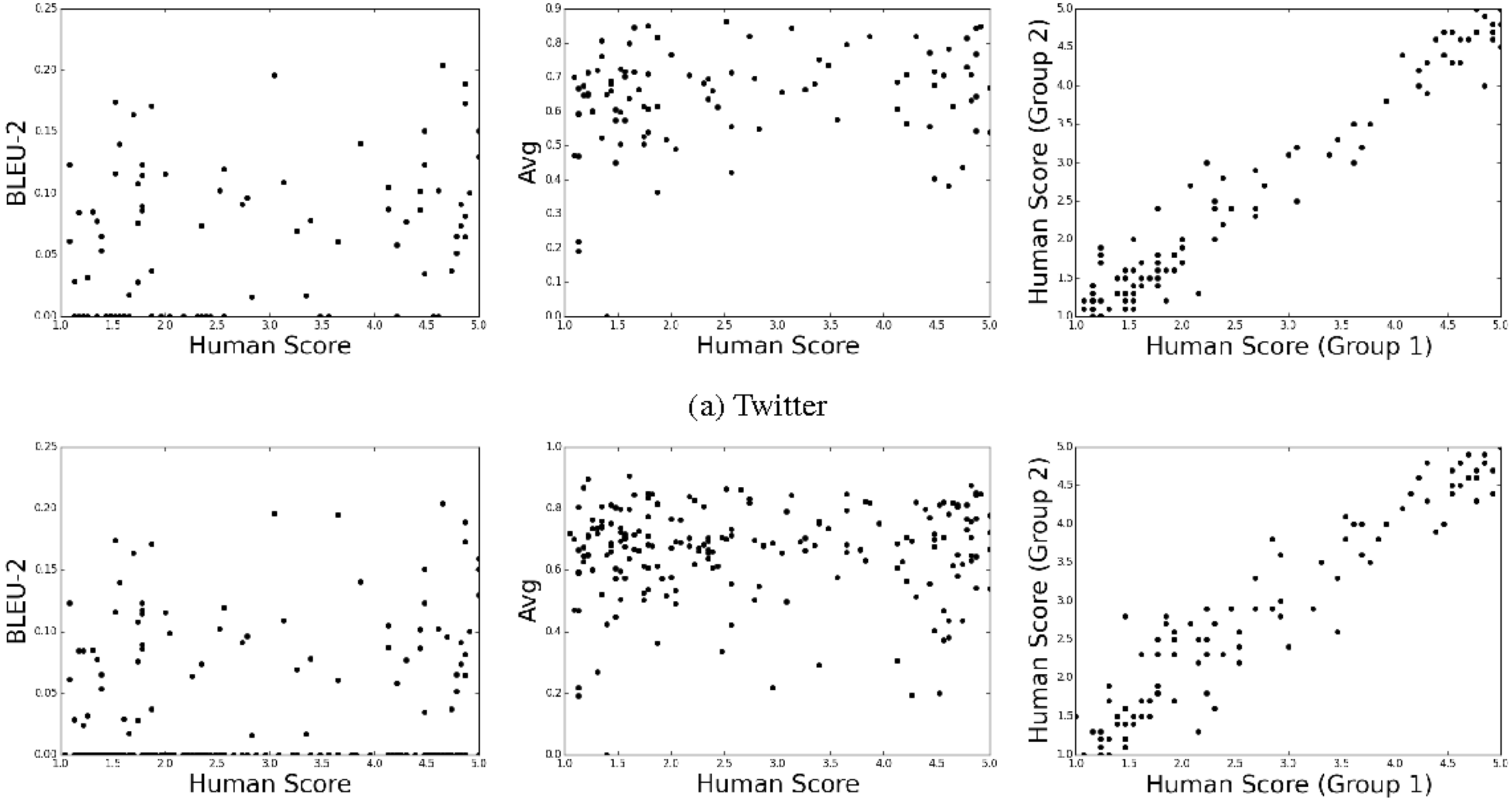
- METEOR
  - Overlap based but uses exact, stem, synonym, and paraphrase matching
    - Denkowski and Lavie, 2014: https://www.cs.cmu.edu/~alavie/METEOR/pdf/meteor-1.5.pdf

Evaluating NLG

- Automated evaluations (word overlap)
  - BLEU
  - ROUGE
  - METEOR
  - F1
- Not great for MT
  - Callison-Burch et al., 2006: https://www.aclweb.org/anthology/E06-1032
  - “Cannot guarantee correlation with humans.”
- Not great for summarization
  - More open-ended than MT
  - Sometimes the highest scoring model has worse output!
    - Paulus et al., 2017: https://arxiv.org/abs/1705.04304
- Not great for dialogue
  - Even more open-ended

Evaluating NLG

- Correlation between metrics and human judges (Liu et al., 2017: https://aclweb.org/anthology/D16-1230)



(a) Twitter

(b) Ubuntu

Evaluating NLG

- Perplexity insufficient
  - Says nothing about actual generated output
- Word embeddings don't capture open ended tasks like dialogue
  - Measures semantic similarity
- Knowledge-based metrics
  - Measure ability to extract knowledge elements from descriptions
    - Whitehead et al., 2018: http://aclweb.org/anthology/D18-1433
    - Lu et al., 2018: http://aclweb.org/anthology/D18-1435
  - Doesn't capture fluency
- Can design metrics specific to what we care about
  - Topical relevance
  - Knowledge conveyed
  - …

Evaluating NLG

- Human evaluation
  - Regarded as gold standard
    - One of our best chances of judging quality output
  - Humans do not get a score of 1.0 BLEU
    - Papineni et al., 2002: https://www.aclweb.org/anthology/P02-1040
  - Human test are very difficult
    - Inconsistent
    - Bias
    - Legal issues
    - …

Outline

- Overview of NLG
  - Summarization
- Image and video captioning
- Knowledge-aware Captioning
- Evaluating NLG

Questions?

Thank you!