Language Comprehension and Language Generation in Eventful Contexts

Nasrin Mostafazadeh

Multi-agent Robotics vs. Natural Language Understanding

Playing Soccer in AI vs. Human

- Multi-agent Learning
- AI Planning
  - Game Strategy Learning
    - Reinforcement Learning
- Motion Planning
- Low-level Control
- Also a complex task for human

Basic NLU in AI vs. Human

Classic Example:

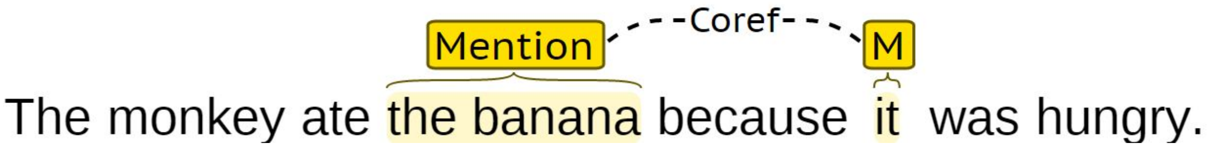
- The monkey ate the banana because it was hungry.
  - What is it? Monkey or the banana?
- Requires enormous amount of knowledge
- *A 4-year old kid can answer this question correctly.*

NLU is Hard

Boston Dynamics' Most Recent Robot

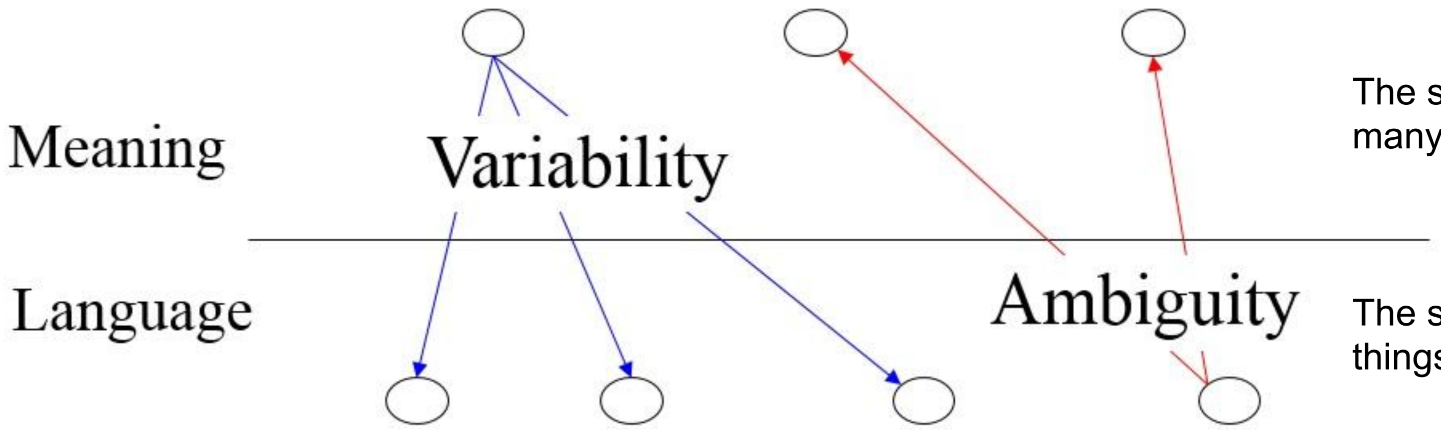
(Feb 2018)

Stanford CoreNLP Coreference Resolver



Why is NLU Hard?

- The Dual Problem of Language Ambiguity and Meaning Variability



The same meaning can be expressed in many ways

The same expression can mean different things and be ambiguous.

Human-level Understanding in Context

Context: At the grocery store

- **Customer**: Black beans?
- **Clerk**: Aisle 3.

Context: Back from the grocery store

- **Woman**: Black beans?
- **Man**: Oh, sorry, forgot to get them.

Context: Serving food

- **Woman**: Black beans?
- **Man**: Yeah, I love it.

Examples Credit: Philip Cohen and James Allen

Context: Serving food

- **Man**: Black beans?
- **Woman**: Oh, you don’t like it?

Fully understanding the underlying linguistic context (no matter how simple) requires the integration of an agent's perception (speech, text, vision, etc.) with its:

- World model
  - Different parties' beliefs and desires
  - The dynamics of events
- Intention Recognition
- Planning
- ...

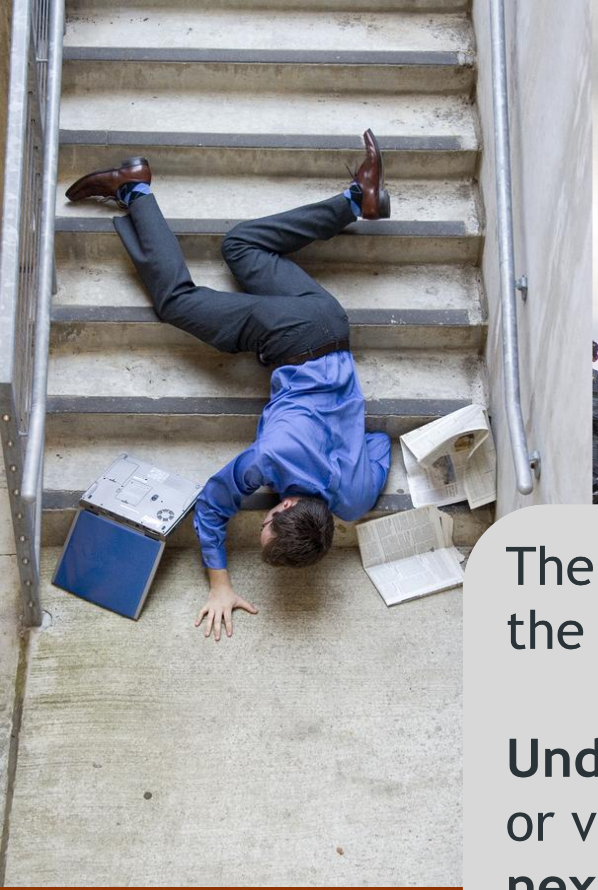
This Talk:

Language comprehension and generation in eventful contexts

With a focus on **commonsense reasoning** and multimodal **context modeling**

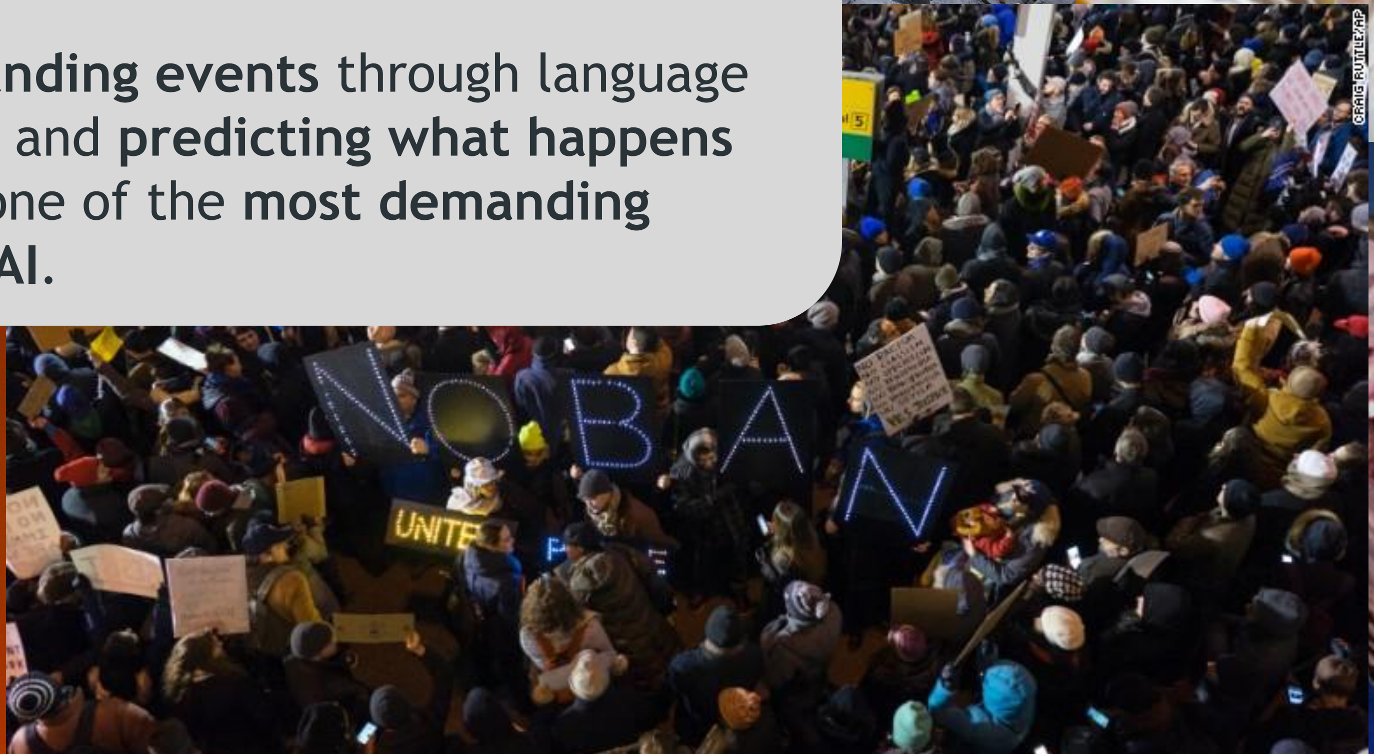
UNIVERSITY of ROCHESTER

Microsoft Research



The changes of the world are **caused** by the **effects** of **events**.

Understanding events through language or vision, and **predicting what happens next**, is one of the **most demanding areas in AI**.



This Talk:

1. **Textual narrative context**
2. Visual context
3. Visual and Textual conversational context
4. Discussion

1. Modeling Textual Narrative Context

Goal: Building a system that can comprehend and collaboratively compose stories with human

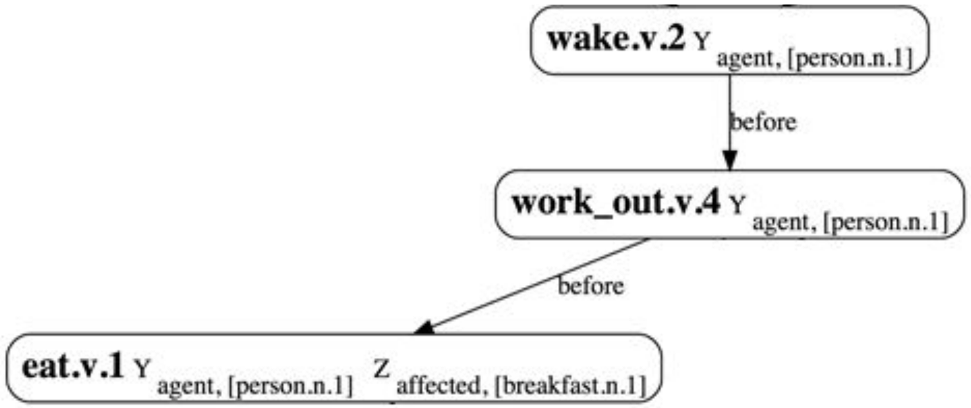
Mostafazadeh et al., NAACL 2016

Story Understanding and Story Generation

- Extremely challenging task in NLP (Charniak 1972; Turner, 1994; Schubert and Hwang, 2000)
- Biggest challenge: **commonsense knowledge** for the interpretation of narratives

How to acquire commonsense knowledge?

- **Scripts (narrative structures):** structured knowledge about stereotypical event sequences together with their participants.



What is a story?

- “A narrative or story is anything which is told in the form of a causally (logically) linked set of events”
  - At this point we are not concerned with how entertaining or dramatic the stories are!

Where to Start Learning Stories/Narrative Structures From?

- We started by machine reading of newswire articles (Chambers et. al., 2008)
  - Not much commonsense knowledge about daily events
- Then, personal stories from blog posts (Gordon et al., 2010)
  - Teasing out useful information from noisy articles was hopeless

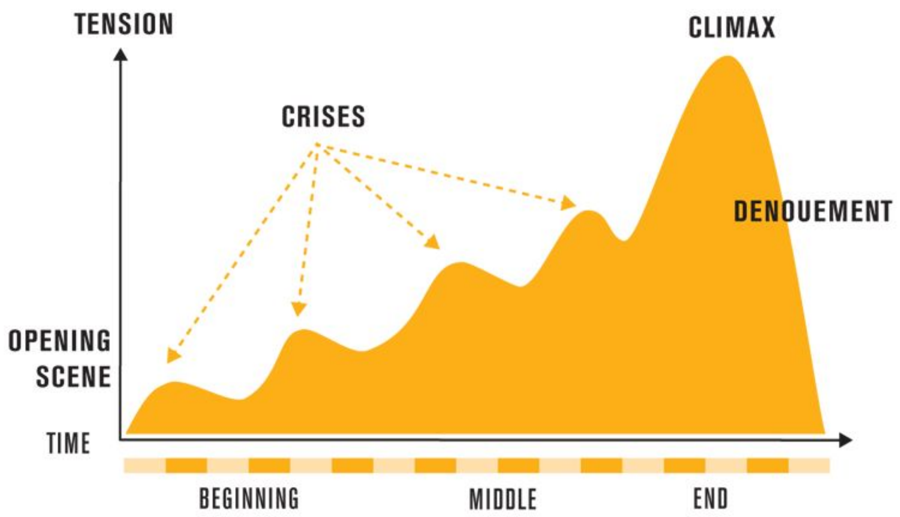
ROCStories

ROCStories: Short Commonsense Stories

- A collection of high quality short **five-sentence stories** with their titles authored by hundreds of **crowd workers**.
    - Enough context to the story, without giving room for sidetracking to less important information

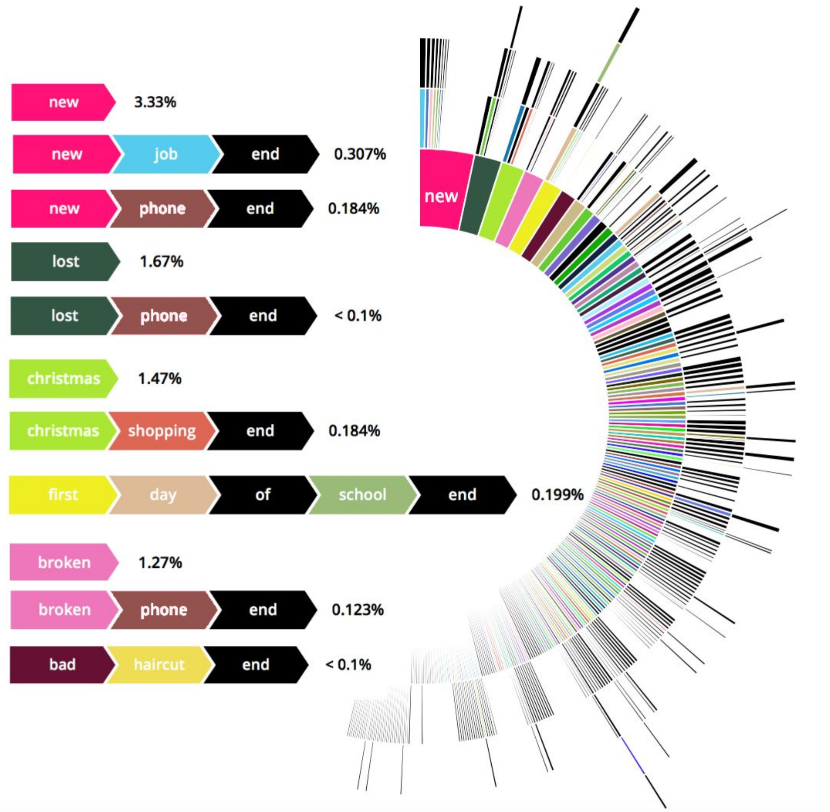
Characteristics:

- Realistic
- Specific beginning and ending, where something happens in between
- Nothing irrelevant or redundant to the core story



Statistics

- **100K** ROCStories
- Total number of Turkers participated: >2000
- Max number of HITs done by one Turker: 4057



An Example Story
Title: “A Friendly Game”

- Bill thought he was a great basketball player. He challenged Sam to a friendly game. He agreed. Sam started to practice really hard. Eventually, Sam beat Bill by 40 points.

X challenges Y —enable→ Y agrees to play —before→ Y practices —before→ Y beats X

Mostafazadeh et al., Event Workshop at NAACL 2016

An Example Story Title: “The President”

- Tom was a great speaker. He talked about hatred and xenophobia in front of large groups of people. People were really inspired by his speech. They decided to vote for him in the election. Tom became the president of the United States.

How to do automatic evaluation on story understanding?

Research has been hindered by the lack of a proper evaluation framework!

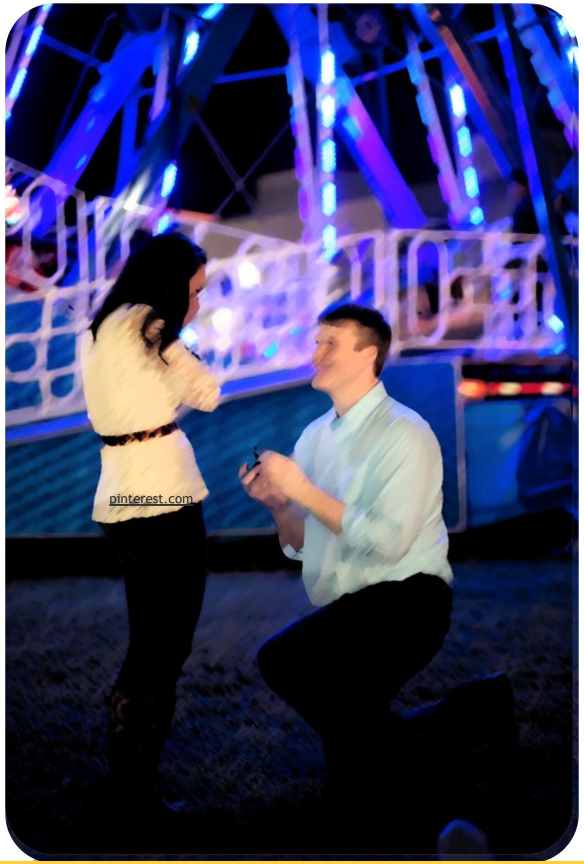
Our Idea: Story Cloze Test (SCT)

- **Goal**: Design a new evaluation schema for story understanding and narrative structure learning.
- **The Story Cloze Test**: Given a context of four sentences, predict the ending of the story.
  - Collect this evaluation dataset of by crowdsourcing

Predicting what happens next

An Example Story Cloze Test

- **Context:** Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down one knee.
- **Right Ending:**
  - Tom asked Sheryl to marry him.
- **Wrong Ending:**
  - He screamed at her and left.



We collected 3,744 **doubly human-verified** Story Cloze Test instances

Story Cloze Models

Learning Typed Narrative Schemes 1/2

Unsupervised model to learn narrative correlation of events

On a large collection of documents

1. Run a dependency parser to extract "event slots"
2. Run coreference resolver to find coreference chains
3. Measure relatedness of each pair of event slots that share an argument
4. Unsupervised clustering of event slots

$$\max_{v \in V} \; narsim(N, v)$$

$$narsim(N,v) = \sum_{d \in D_v} \max(\beta, \max_{c \in C} chainsim(c, \langle v,d \rangle))$$

$$chainsim(c, \langle v,d \rangle) = \max_{a \in Args} (score(c,a) + \sum_{i=1}^{n} sim(\langle e,d \rangle, \langle v,d \rangle, a))$$

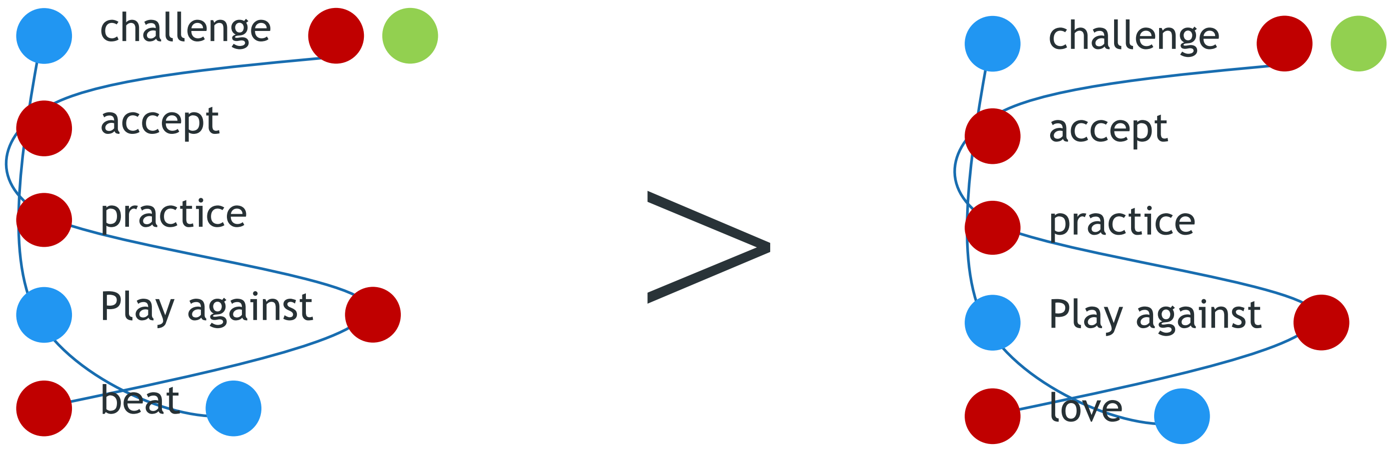
$$sim(\langle e,d \rangle, \langle v,d \rangle, a) = pmi(\langle e,d \rangle, \langle v,d \rangle) + \lambda \log C(\langle e,d \rangle, \langle v,d \rangle, a)$$

Chambers & Jurafsky, ACL 2009

Learning Typed Narrative Schemes 2/2

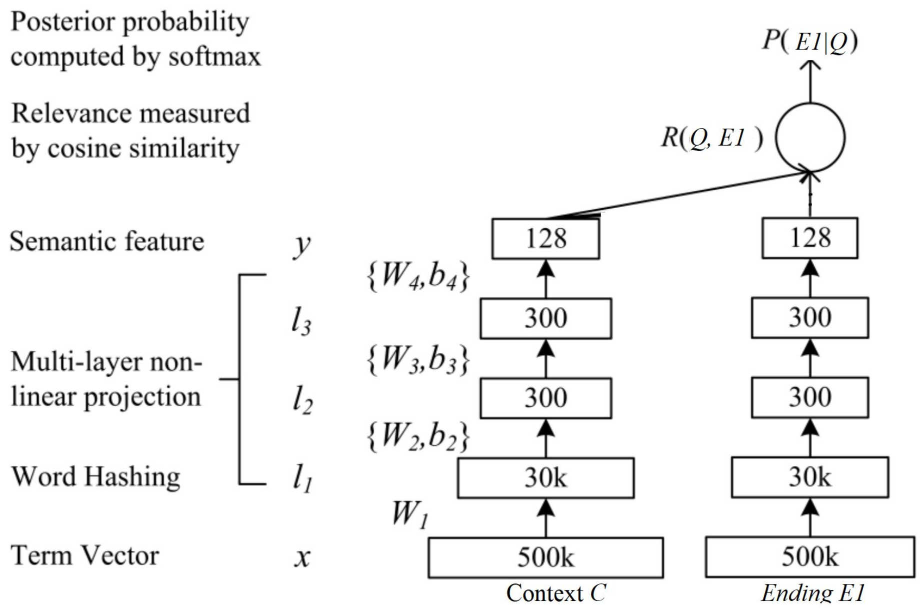
At test time

- Choose the ending which yields the higher total *narsim(N)* for the resulting narrative structure *N*



Deep Structured Semantic Model

- **Deep Structured Semantic Model (DSSM)**
  - Sentence2Vec model (Huang et al., CIKM 2013), trained two letter-n-gram NNs to project the four-sentences context and the fifth sentence into the same vector space, so that the right ending has the smaller cosine distance.



Baseline Models

- **Frequency** (discard the context): Choose the ending with higher (search engine hits) frequency of the main event.
- **N-gram overlap**: Choose the ending with higher n-gram overlap with the context, computed using Smoothed-BLEU metric.
- **Average Word2Vec (neural BOW):** Choose the ending with closer average word2vec to the average word2vec of the four-sentences context.
- **Sentiment Match**: Choose the ending that matches the sentiment of the four-sentences context (Full) or the fourth-sentence (Last).
- **Skip-thoughts Model**: Toronto's Sentence2Vec encoder which models the semantic space of novels (stories), according to which you can choose the option that has a closer embedding to the four-sentences context.

Results

- Accuracy = $\frac{\#\ correct\ choices}{\#\ test\ cases}$

	Constant-choose-first	Frequency	N-gram-overlap	GenSim	Sentiment-Full	Sentiment-Last	Skip-thoughts	Narrative-Chains-AP	Narrative-Chains-Stories	DSSM	Human
Validation Set	0.514	0.506	0.477	0.545	0.489	0.514	0.536	0.472	0.510	0.614	1.0
Test Set	0.513	0.520	0.494	0.539	0.492	0.522	0.552	0.478	0.494	**0.595**	1.0

Story Cloze Test

The benchmark for narrative understanding

- Human performs 100%
- A challenging task with a wide enough gap (42%) from the state-of-the-art and human performance, so plenty of room for

- Various use-cases
  - Training models which understand or tell stories
  - Training generic language models
  - Evaluating children’s intellectual disabilities!
  - Developing theories of what makes a sequence a story.
  - …
- List of all papers and resources related to ROCStories project http://cs.rochester.edu/nlp/rocstories/

Story Cloze Shared Task

- Time was ripe to organize the first SCT challenge
  - LSDSem EACL workshop
    - **18 teams registered to participate**
    - **8 teams participated**
    - Used the original Story Cloze Test Set – Spring 2016 for evaluation
- A variety of submitted approaches
  - Rule-based methods
  - Linear classifiers using different discourse phenomena
  - End-to-end neural models
  - Hybrid models

Use DNN in some way

Results

#	User	Percent
1	msap	0.752004 (1)
2	cogcomp	[illegible]
3	tbmihaylov	0.724212 (3)
4	ukp	[illegible]
5	Niko	[illegible]
6	roemmele	0.671833 (6)
7	mflor	0.620524 (7)
8	Pranav_Goel	0.604490 (8)
9	ROCNLP	0.595938 (9)
10	lizhongyang	0.585249 (10)
11	sjtuadapt	0.585249 (10)

Use Pre-trained Embeddings

Report on ‘sentiment’ being an important factor

Current SOTA, UIUC team

Story Comprehension for Predicting What Happens Next

EMNLP'17

Results

#	User	PercentageScore ▲
1	cogcomp	0.776056 (1)

Hey, Juicero!


JUICERO
CHOPPED
Greens
FRESH CHOPPED PRODUCE FOR MAKING COLD-PRESSED JUICE
JUICERO

Beautiful Engineering

& Our Obsession with Complexity …

Our Love for Model Complexity… 1/2

Learning Typed Narrative Schemes 2/3

- person, person, game

$$\max_{v \in V} \; narsim(N,v)$$

$$narsim(N,v) = \sum_{d \in D_v} \max(\beta, \; \max_{c \in C} chainsim(c,\langle v,d \rangle))$$

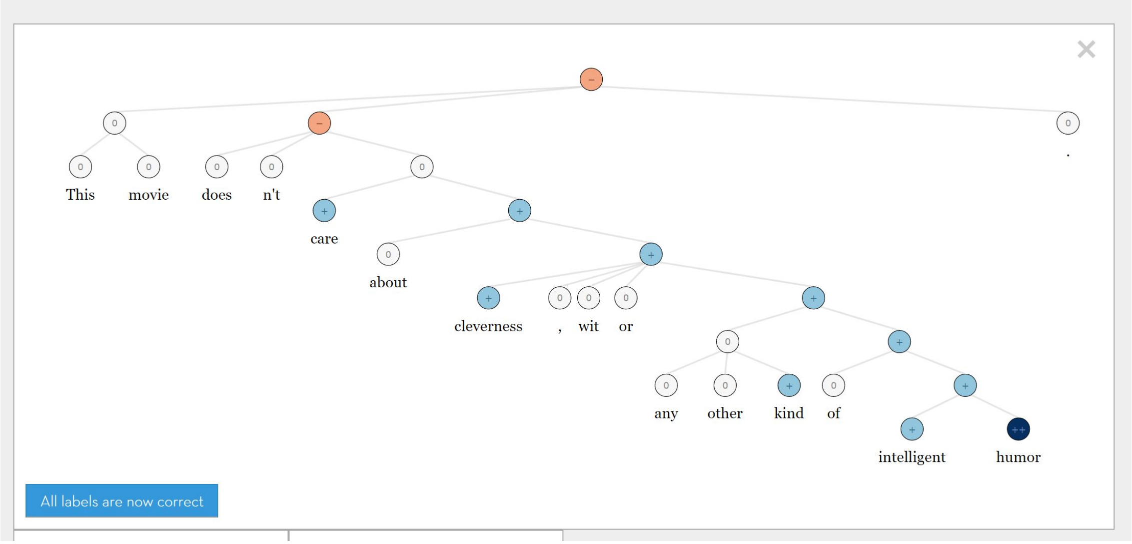
$$chainsim(c,\langle v,d \rangle) = \max_{a \in Args} (score(c,a) + \sum_{i=1}^{n} sim(\langle e,d \rangle,\langle v,d \rangle,a))$$

$$sim(\langle e,d \rangle,\langle v,d \rangle,a) = pmi(\langle e,d \rangle,\langle v,d \rangle) + \lambda \log C(\langle e,d \rangle,\langle v,d \rangle,a)$$

Chambers & Jurafsky, ACL 2009

- Romelle et al. (2017) computed basic PMI score for all the word pairs of context: achieve **59.9** vs **49.4**

Our Love for Model Complexity … 2/2



We used “Recursive Neural Network” sentiment analyzer trained on ~12,000 sentences and achieved **49.2**

- Goel & Singh (2017) : Use VADER (a rule-based sentiment analyzer) for sentiment-match and achieve **58.2**

What’s next for the Story Cloze Test?

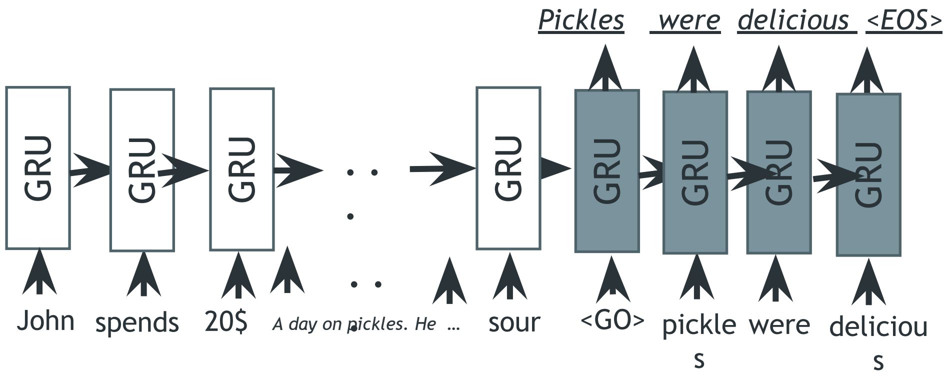
- We are very encouraged by the level of participation in the first shared task!
- There is **still a large gap (23%) between the current SOTA** and **Human performance** even on **the current test set.**
- We have implemented some new crowdsourcing and human verification steps for isolating any possible data collection/writing style artifacts
- Updates via http://cs.rochester.edu/nlp/rocstories/

Collaborative Story Composition

Given the narrative context from human, the system should generate the ending, as opposed to classification

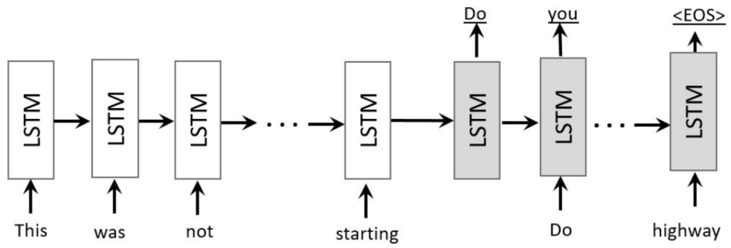
Sequence2Sequence generation
(Sutskever et. al, 2014)

- Encoder-Decoder RNN Architecture with Attention
- 2-layers, with 512 units per layer
- Beam-search decoding, with beam-width = 25



Fails when feeding in the four sentence context to generate the ending

Context = 4th sentence



- **4th Sentence:** She wrote another through self-publishing avenues.
- **System:** And they laughed !

- **4th Sentence:** She got it painted a bright pink color.
- **System:** Kelly is very happy .

- **4th Sentence:** He did extra credit assignments.
- **System:** Everyone loved it .

- **4th Sentence:** At midnight, he looked back and saw Anita dancing with another guy.
- **System:** He was thrilled when he was the man he wanted to be !

- It is easy to generate something that kind of makes sense … when you **discard the bigger context**!
- As humans, we have the amazing capability of imagining a scenario for validating even a random output whenever there is no/short established prior context

The importance of Modeling (non-trivial) Context

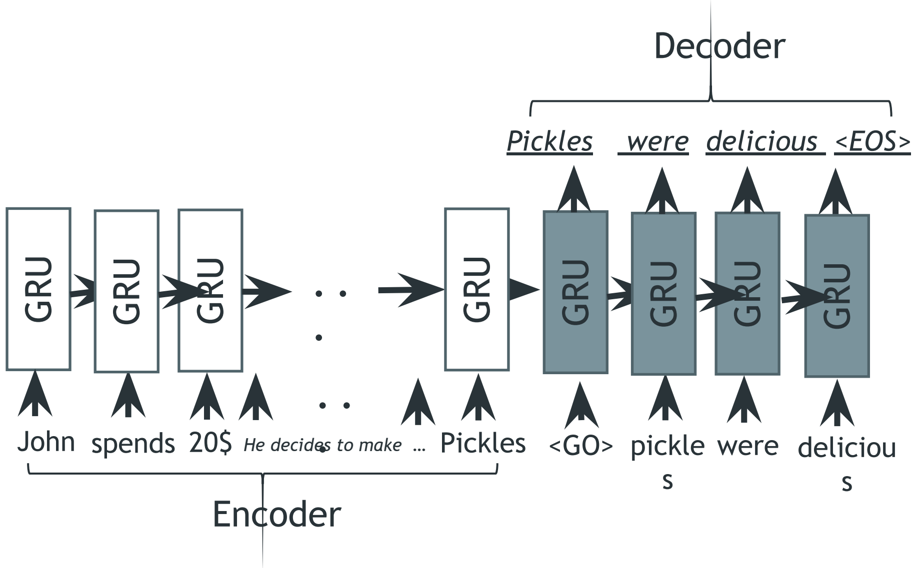
Modifying the context: only retaining main events

Input 4-sentence story (only highlighted words)**:**

John spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. John waits 2 weeks for his pickles to get sour.

Output fifth-sentence:
Pickles were delicious <EOS>

<EOS>= end of sentence marker
- We extract the **main events and their semantic roles** using TRIPS STEP parser.



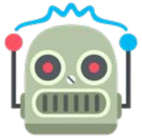
Example Generation

- **Context**: John spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. John waits 2 weeks for his pickles to get sour.

- **1:** Now he is so happy that he has money.
- **2:** John opens the jar to find perfect pickles.

Example Generation

- **Context**: Kara lost her dog yesterday. She freaked out. Frantic, Kara looked all over town for it. Finally she found the dog in her neighbor's yard.

- **1:** She was relieved to have it back home. 
- **2:** She asked the neighbors not to put out treats. 

But More Often …

- **Context:** Ivy wanted a cell phone like all her friends. She saved up her money from delivering papers. Soon she had enough for a phone. She went shopping and selected the perfect <u>phone</u>.
- **System**: She was so happy she was gotten her food !

- **Context:** Nicole wanted to be an Aunt more than anything. The trouble was that she didn't have any siblings. Today Nicole got a note from her cousin, Xiao. Xiao and her husband were expecting a baby boy.
- **System**: He was glad he did not .

Better Context Representation

- John spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. John waits 2 weeks for his pickles to get sour.

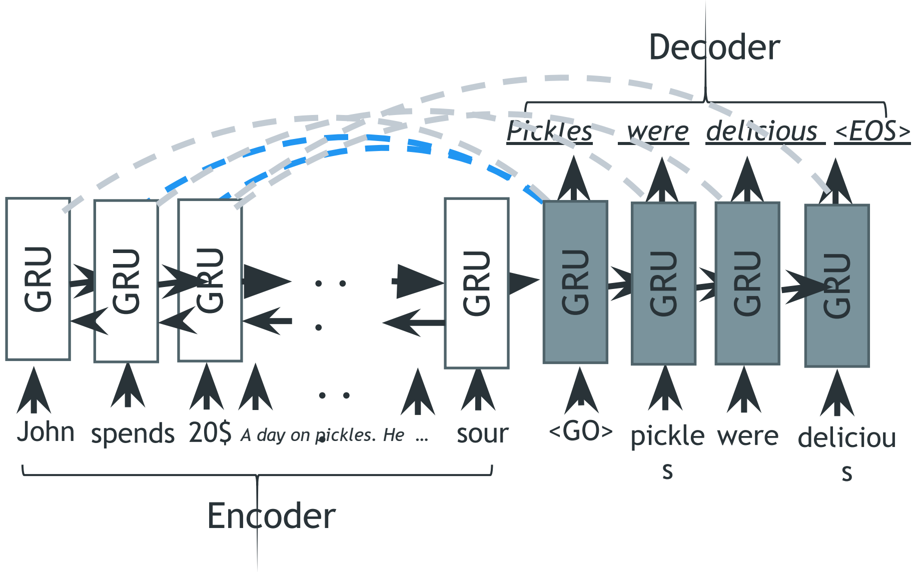
Better Context Representation

- Preprocessing:
  - NER
  - Coreference Resolution
  - Abstraction using Ontology Type

- **PERSON1** ONT::commerce-pay $20 a day on ONT::condiment. **PERSON1** ONT::decide to ONT::create **PERSON1*** to ONT::save-cost ONT::money. **PERSON1** puts the ONT::condiment in ONT::brine. **PERSON1** ONT::waits **DURATION1** for **PERSON1*** ONT:condiment to ONT:become ONT:sour.

Sequence2Sequence Generation

- Bi-Directional Encoder-Decoder RNN Architecture with Attention
- 2-layers, with 512 units per layer
- Beam-search decoding, with beam-width = 25
- Reranking using PRO algorithm



Chris Manning's BiLSTM (with attention) Hegemony!

- Trained on 400K (story context, next utterance) pairs

Collaborative Turn-by-Turn Generation

- **PERSON1** ONT::commerce-pay $20 a day on ONT::condiment.

😎

- **PERSON1** decided to go to the store.

🤖

- **PERSON1** ONT::purchase more ONT:condiment.

😎

- **PERSON1** was very happy.

🤖

Generate the Ending

- **PERSON1** ONT::commerce-pay $20 a day on ONT::condiment. **PERSON1** ONT::decide to ONT::create **PERSON1*** to ONT::save-cost ONT::money. **PERSON1** puts the ONT::condiment in ONT::brine. **PERSON1** ONT::waits **DURATION1** for **PERSON1*** ONT:condiment to ONT:become ONT:sour.😎
- **PERSON1** was very proud.🤖

Language Generation

Where are we standing?

- RNNLMs are performing great on generating grammatical outputs
  - Local coherency
- **Logically-sound** generation, given context, is still very challenging
  - Generation given a trivial context (a topic, a title, or a sentence) is easier than generating a **logically-sound output** given **an established non-trivial long context**
  - Since as humans we are great at hypothesizing scenarios for rationalizing almost any random sequence without an established context!
- Generating Shakespeare-like text, poetry, or fictitious text is not as challenging
  - Since often irrelevant content can be also deemed “creative” by human!

What is still very hard?
“to generate a **contentful** sequence of **logically related** sentences.”

Better Narrative Context Representation

Ongoing Work

- We need models that learn to ‘generalize’ better
  - Any training corpus for a generation task requiring commonsense knowledge will be small, if we don’t work on better ‘**abstraction**’
  - We should leverage semantic abstractions for better context representation

2. Modeling Visual Context

Goal: Building a system that can ask a natural question given an eventful image as the context

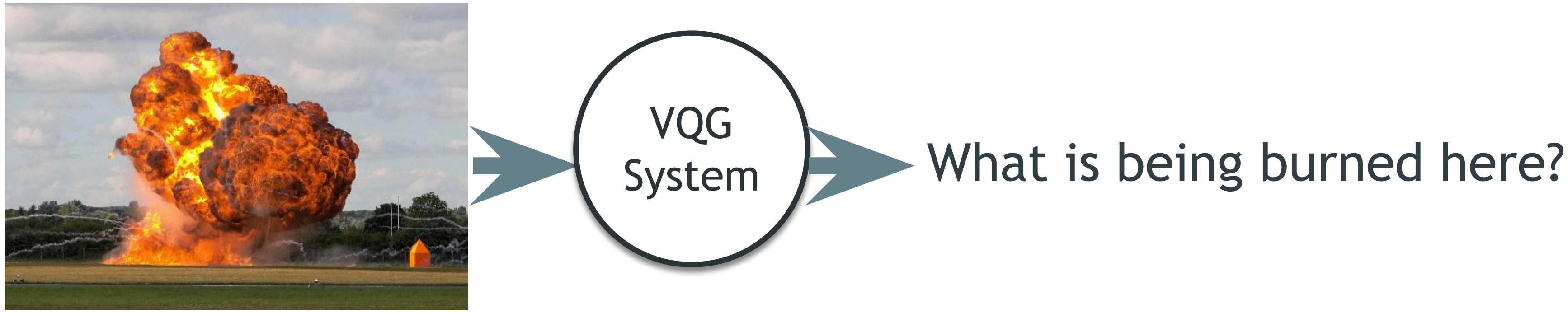
Mostafazadeh et al., ACL 2016

What is the very first question that comes to your mind?

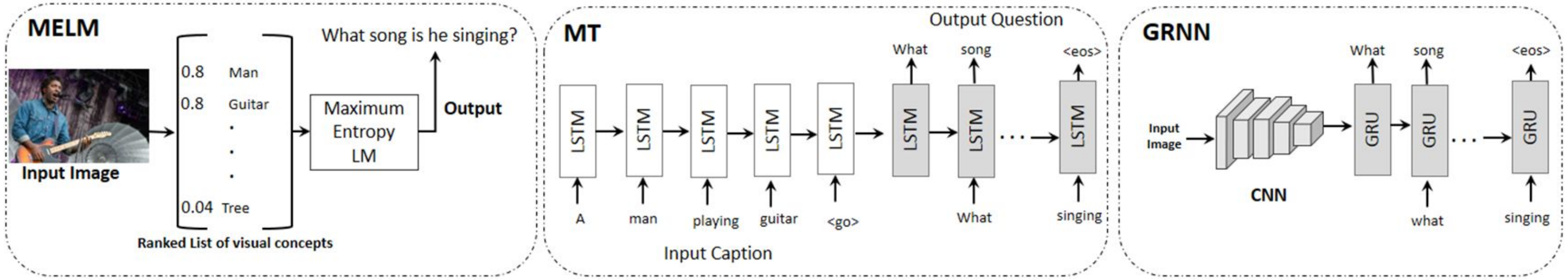
Visual Question Generation (VQG)

- We introduced the task of VQG
  - Asking the 'right' question shows intelligence
- To enable this task, we crowdsourced three VQG datasets from various resources, from **object to event-centric**, each with 5,000 images and 5 questions per image:
  - VQG_{COCO}
  - $\text{VQG}_{\text{Flickr}}$
  - VQG_{Bing} Queried Bing with **event-centric** keywords

Models



Generation Models



- Sequence-to-sequence GRNN Captioning (Devlin et al. 2015; Vinyals et al., 2015, Cho et al., 2014)
  - Transform the *fc7* vector to 500 dimensional vector to be initial state to a 500-dimensional Gated Recurrent Unit (GRU)
  - Vocabulary size 1942 tokens
  - Constrained to generate questions >6 tokens

Example Generation

- What caused the damage to this city?
- **GRNN**: What happened to the city?
- **KNN**: What state was this earthquake in?
- **Caption Bot**: A pile of dirt.

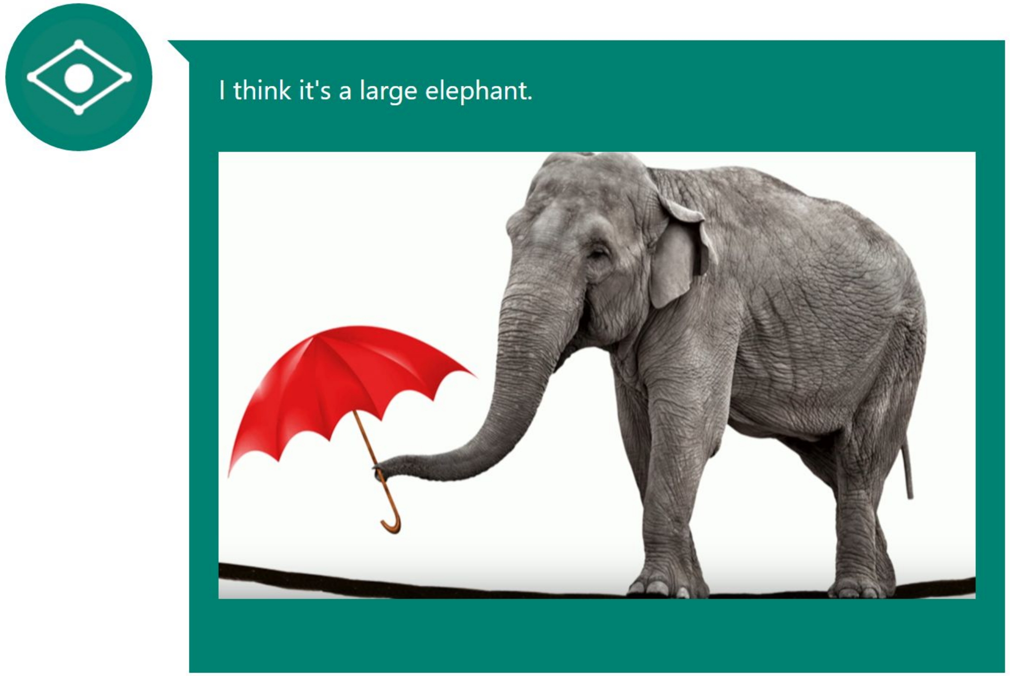
Example Generation

- Did the drivers of this accident live through it?
- **GRNN**: How did the car crash?
- **KNN**: Was anybody hurt in this accident?
- **Caption Bot**: A man standing next to a motorcycle.

Image Captioning

Out of the scope of the training data

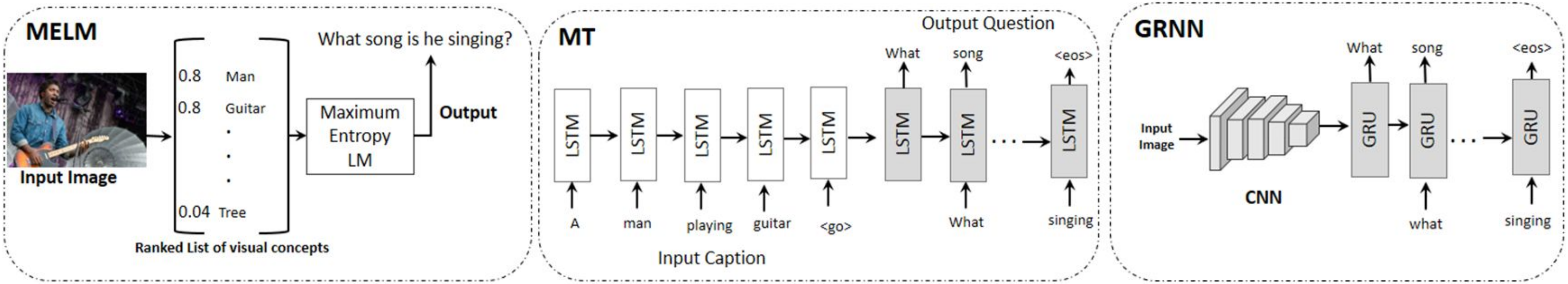
BLEU		METEOR	
Bing	*MS COCO*	*Bing*	*MS COCO*
0.101	0.291	0.151	0.247



Visual Question Generation

Out of the scope of the training data

Generation Models & Results



	$Human_{consensus}$	$Human_{random}$	$GRNN_X$	$GRNN_{all}$	$1\text{-}NN_{bleu-X}$	$1\text{-}NN_{gensim-X}$	$K\text{-}NN\text{+}min_{bleu-X}$	$K\text{-}NN\text{+}min_{gensim-X}$	$1\text{-}NN_{bleu-all}$	$1\text{-}NN_{gensim-all}$	$K\text{-}NN\text{+}min_{bleu-all}$	$K\text{-}NN\text{+}min_{gensim-all}$
Human Evaluation												
Bing	2.49	2.38	1.35	**1.76**	1.72	1.72	1.69	1.57	1.72	1.73	1.75	1.58
COCO	2.49	2.38	1.66	1.94	1.81	1.82	1.88	1.64	1.82	1.82	**1.96**	1.74
Flickr	2.34	2.26	1.24	**1.57**	1.44	1.44	1.54	1.28	1.46	1.46	1.52	1.30

	METEOR	**BLEU**	**ΔBLEU**
r	**0.916** (4.8e-27)	0.915 (4.6e-27)	0.915 (5.8e-27)
ρ	0.628 (1.5e-08)	0.67 (7.0e-10)	**0.702** (5.0e-11)
τ	0.476 (1.6e-08)	0.51 (7.9e-10)	**0.557** (3.5e-11)

3. Modeling Visual & Textual Context

Goal: Building a system that can engage in a natural conversation about an eventful image

Mostafazadeh et al., IJCNLP 2017

😎Yes he won, he can’t believe it.

Did he end up winning the race?

Image-Grounded Conversations

😎 My son is ahead and surprised!

🤖 Did he end up winning the game?

😎 Yes he won, he can't believe it.

Visual/situational context

Discourse Context

Proactively drive the conversation forward by asking "reasonable" questions!

Image-Grounded Conversations (IGC)

- IGC is on the continuum between chit-chat models of conversation, and the goal-directed conversation systems.
  - Visually grounding conversations in an eventful image **naturally serves to constrain the topic of conversation.**
- We focus on questions as conversation openers!

Image-Grounded Conversations

Twitter Data Example

I found a cawaii bird.

Are you going to collect some feathers?

There are so many crows here I'd be surprised if I never found one.

Image-Grounded Conversations on Eventful Images

Crowd

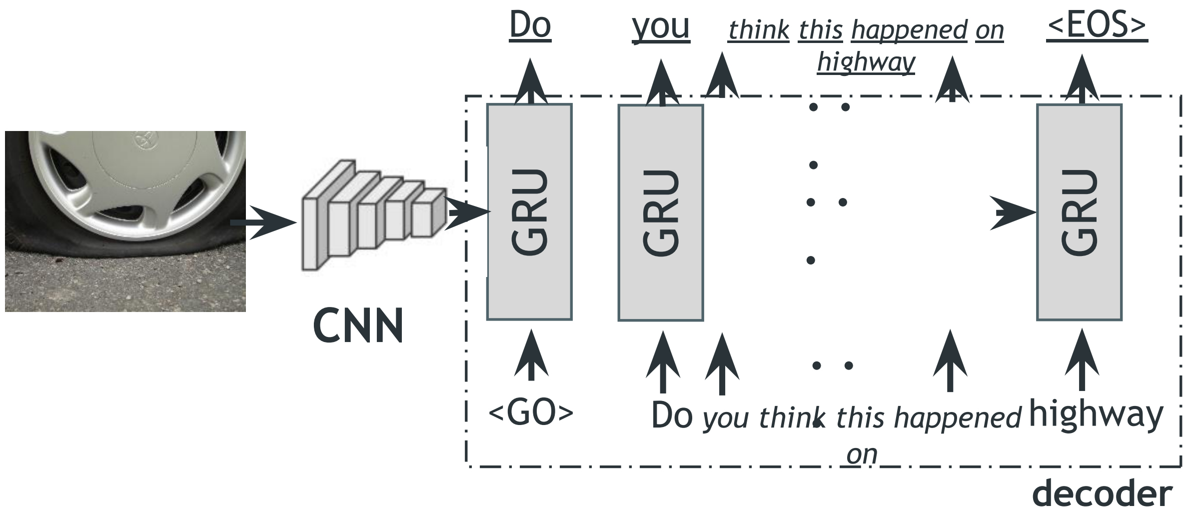
A terrible storm destroyed my house!

OH NO, what are you going to do?

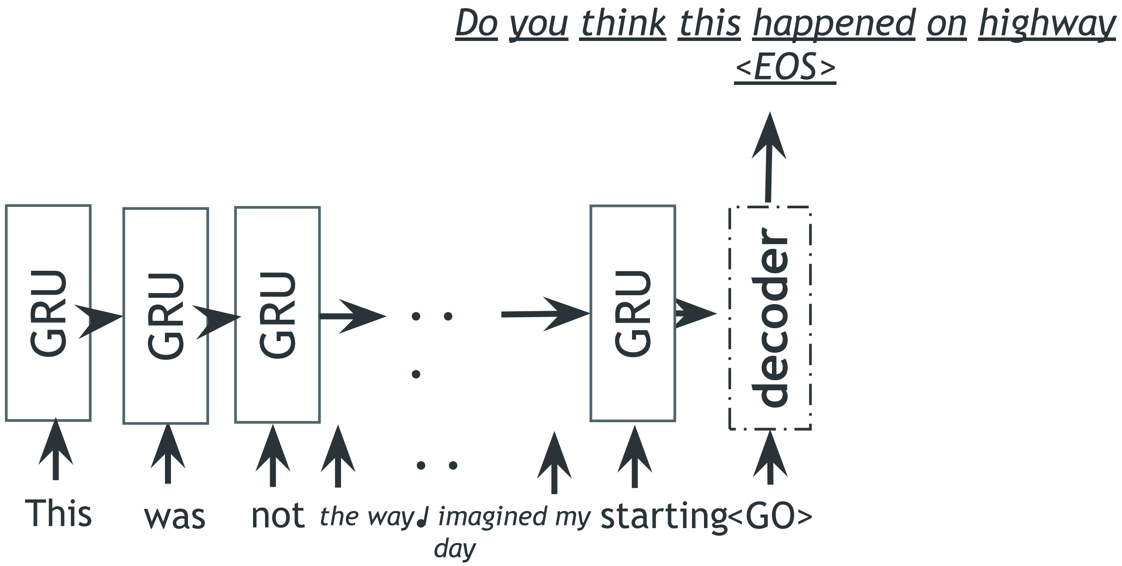
I will go live with my Dad until the insurance company sorts it out.

Models

Visual Context Sensitive

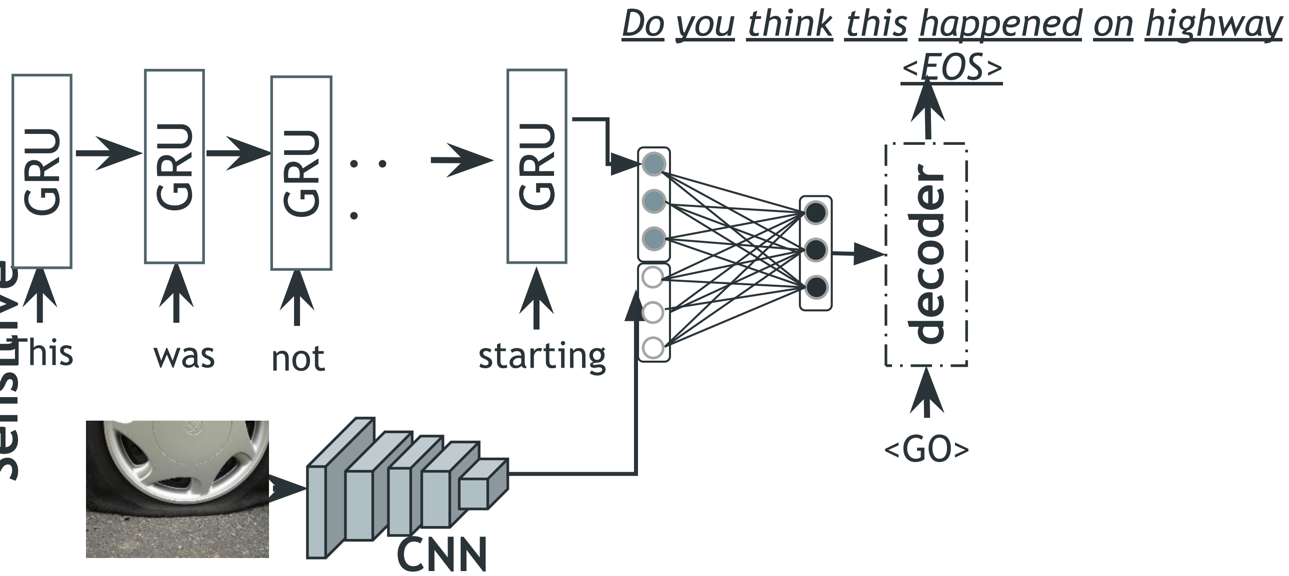


Textual Context Sensitive



Models

Visual & Textual (RNN) Context Sensitive



Visual & Textual (BOW) Context Sensitive

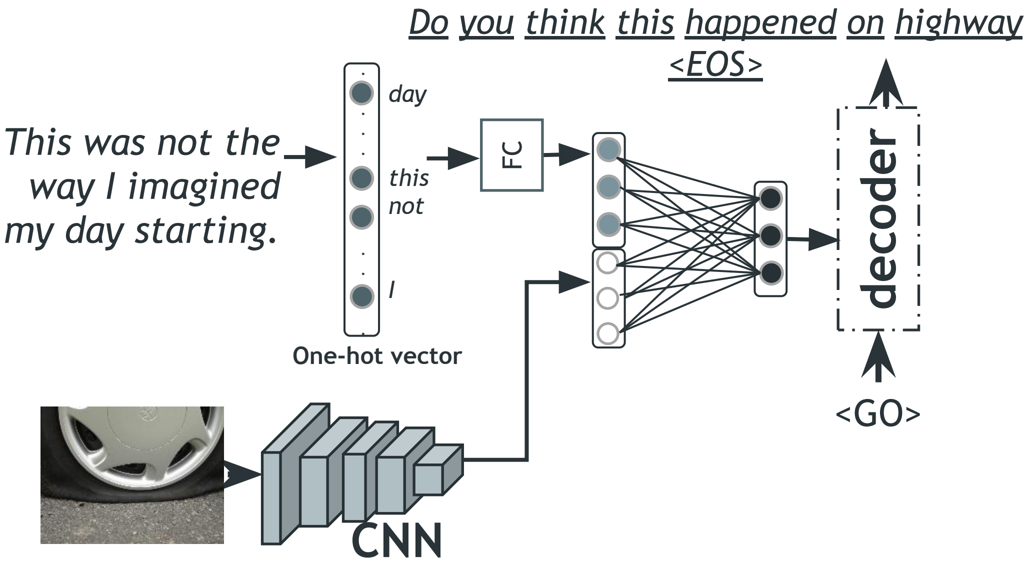


Image-Grounded Conversations

Question Generation

I got in a car wreck today!

What happened?

Well, drunk driving!

Image-Grounded Conversations
Response Generation

I got in a car wreck today!

Did you get hurt?

Nah, I’m home now!

Image-Grounded Conversations
Response Generation

I got in a car wreck today!

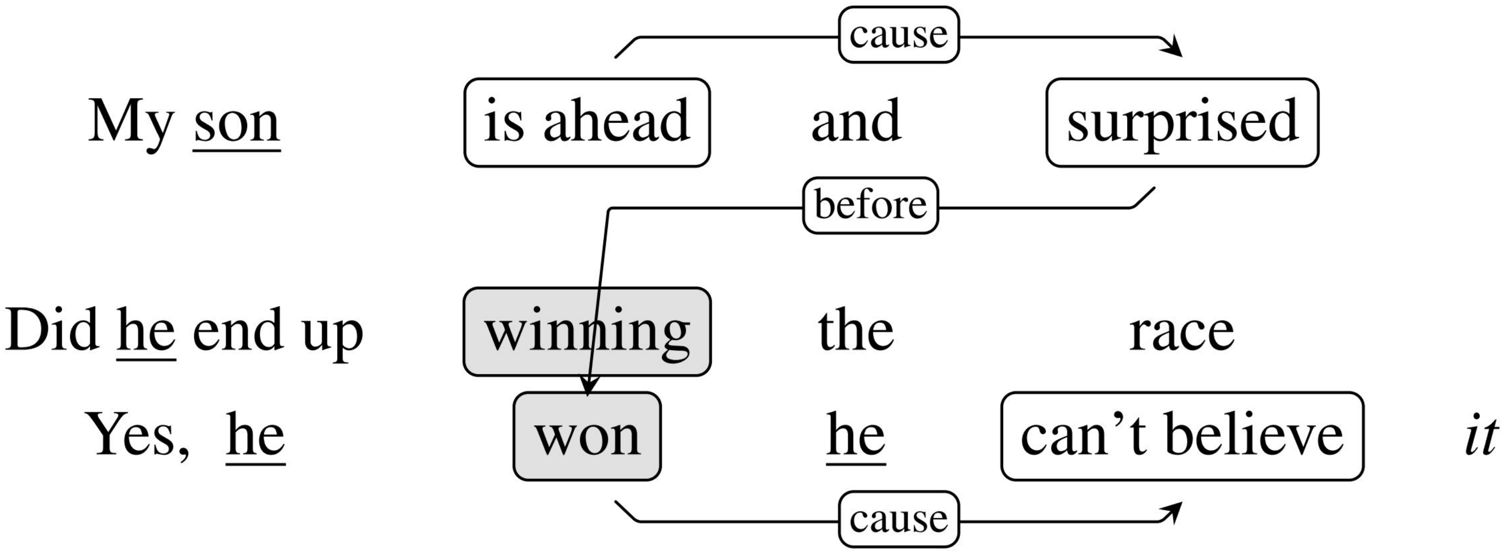
Did you get hurt?

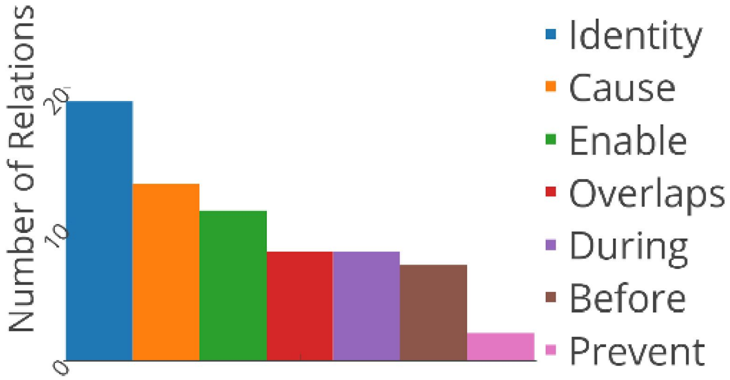
Nah, I'm home now!

Story:
Sam got in a car wreck today. He did not get hurt. He managed to get home …

Causal and Temporal Relation Scheme (CaTeRS) in Eventful Grounded Conversations

Mostafazadeh et. al, Event Workshop at NAACL 2016

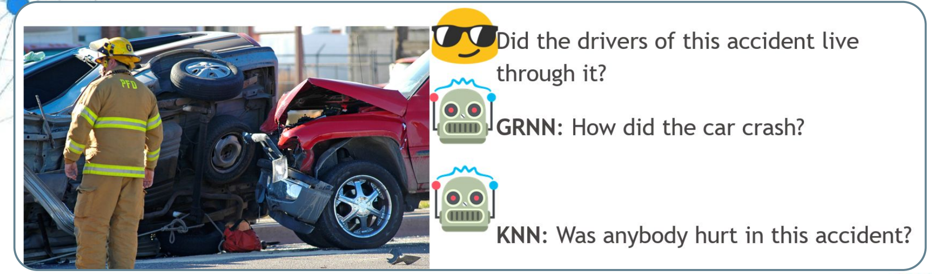




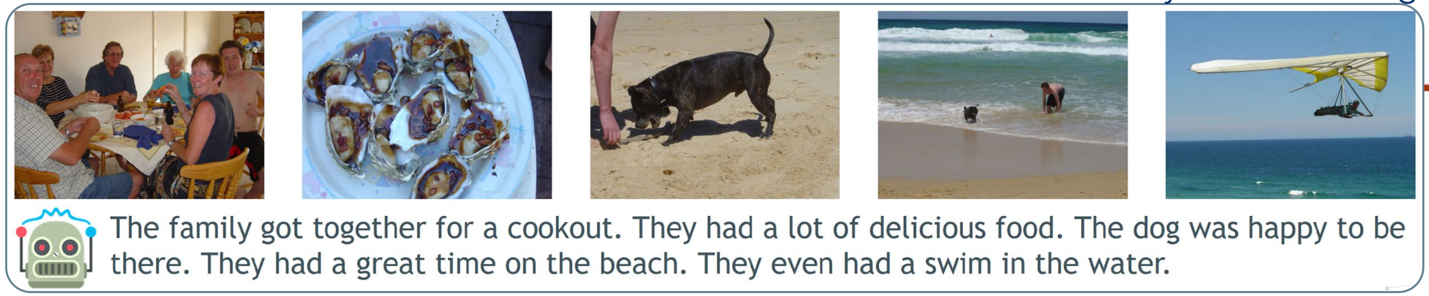
Human Evaluation on Question & Response Generation

		Human	Generation (Greedy)			Generation (Beam, best)			
		Gold	Textual	Visual	V & T	Textual	Visual	V & T	VQG
Q	Crowd	2.68	1.46	1.58	1.86	1.07	1.86	**2.28**	2.24
R	Crowd	2.75	1.24	–	1.40	1.12	–	**1.49**	–

Discussion



Visual Question Generation



Visual Storytelling

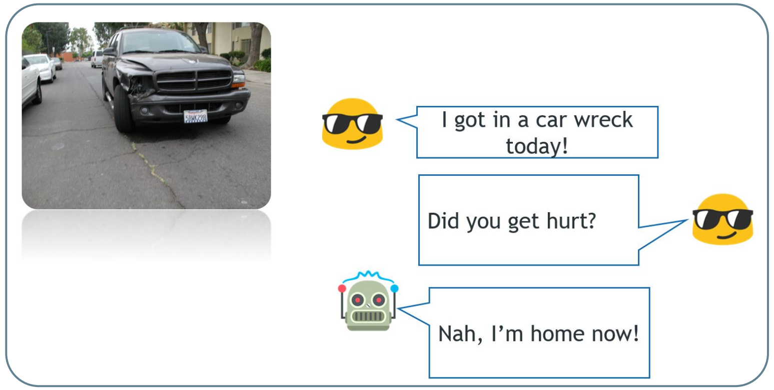
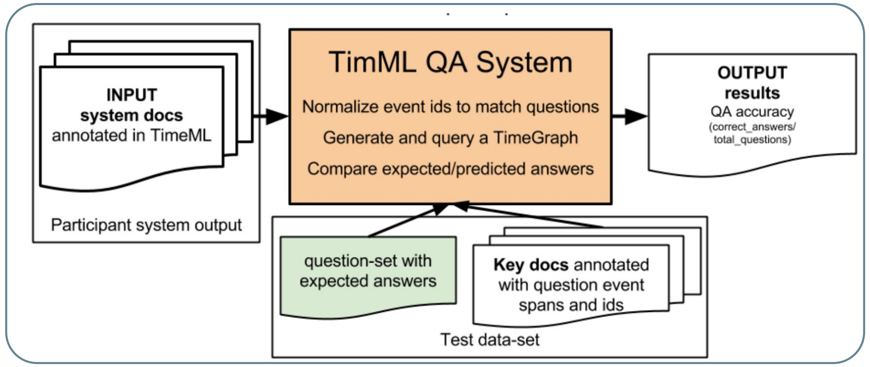


Image-Grounded Conversations



Temporal Question Answering

Context: John spends $20 a day on pickles. He decides to make his own to save money. He puts the pickles in brine. John waits 2 weeks for his pickles to get sour.

1: Now he is so happy that he has money.

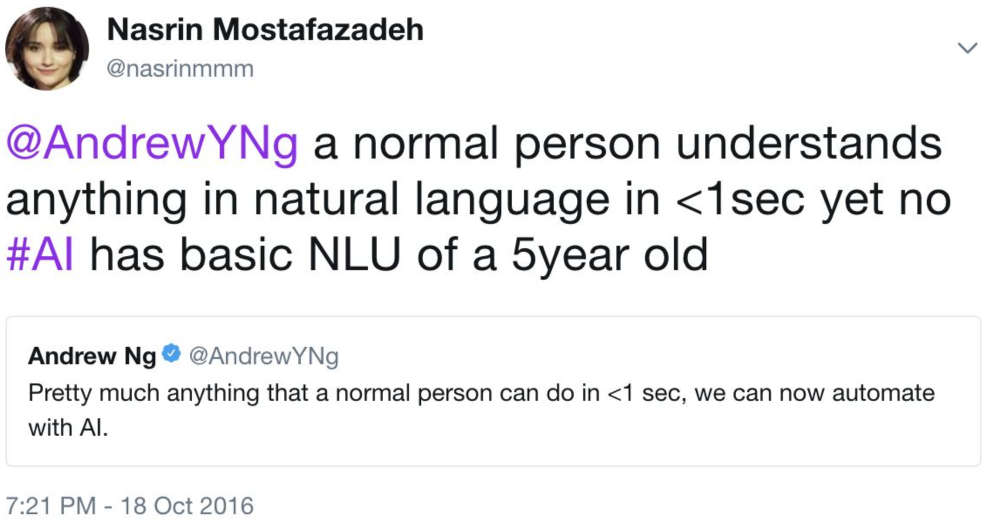
Story Cloze Test & Story Generation

The Current Trend in AI and NLP

- For a particular narrow task:
  - Build a large dataset
    - Scalable via crowdsourcing
  - Design a complex model
    - May or may not establish “strong” baselines
  - Use the dataset to train and test the new model
    - In practice, end up finding correlations and patterns in data and often overfit to the intricacies and biases of the dataset
    - Often fail at real-world non-biased test cases
- Repeat for a new task!

What's Lacking?

- We've made a great progress in perception tasks such as 'speech recognition' and 'image recognition'
- There is a consensus that "commonsense reasoning remains fundamentally unsolved today in AI"
  - The AI community has to move towards **reasoning**



What can we do?

- Move away from task-specific trained models and annotated datasets

  i. Fully supervised models are not applicable when collecting a large annotated dataset is infeasible

    - We need better ways of **abstraction** and **generalization** to **transfer knowledge** from a task to another

    - We should consider **various supervision scenarios**

  ii. Ground predictions in a more **complex** and **realistic contexts**

    - Reasonable benchmarks with strong baselines

    - Contentful contexts are often **event-centric**

What can we do?

- Move away from task-specific trained models and annotated datasets
  1. Fully supervised models are not applicable when collecting a large annotated dataset is infeasible
     - We need better ways of **abstraction** and **generalization** to **transfer knowledge** from a task to another
     - We should consider **various supervision scenarios**
  2. Ground predictions in a more **complex** and **realistic contexts**
     - Reasonable benchmarks with strong baselines
     - Contentful contexts are often **event-centric**
- **+ Move towards tangible applications in real world!**

Thanks to

James Allen, Lucy Vanderwende, Nathanael Chambers, Pushmeet Kohli, Margaret Mitchell, Chris Brockett, Bill Dolan, Michel Galley, Xiaodong He, Devi Parikh, Dhruv Batra, Ishan Misra, Jacob Devlin, Jianfeng Gao, Alyson Grealish, Rishi Sharma

Thanks for Listening

Any Questions?