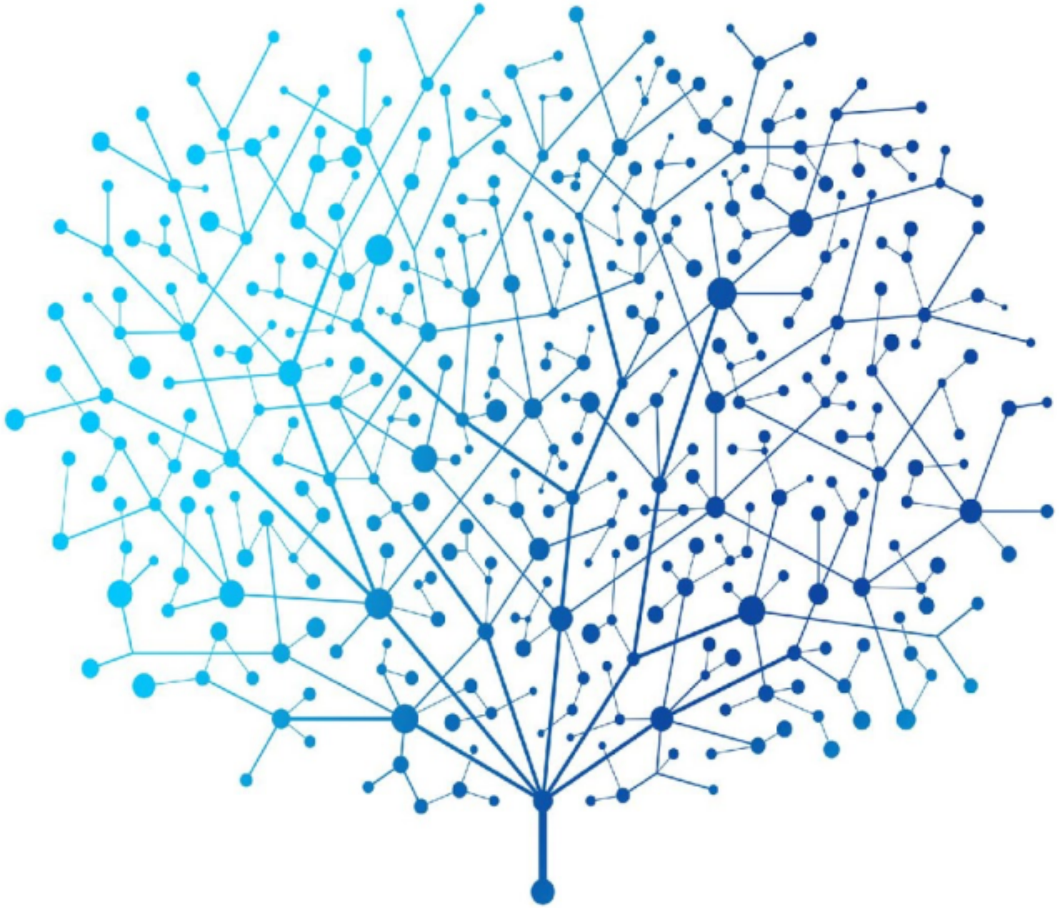
Language Comprehension and Language Generation in Eventful Contexts

Nasrin Mostafazadeh

BenevolentAI



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.

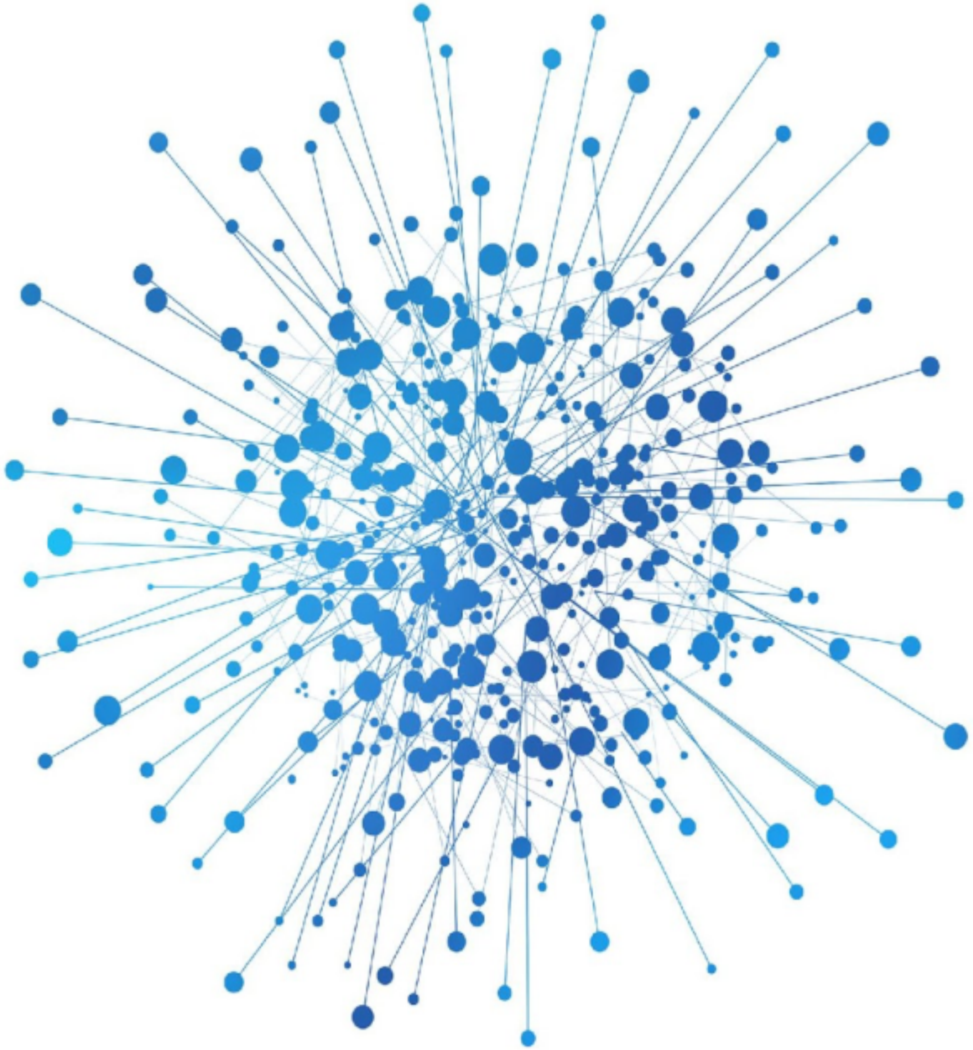
Examples Credit: Philip Cohen and James Allen

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
- ...

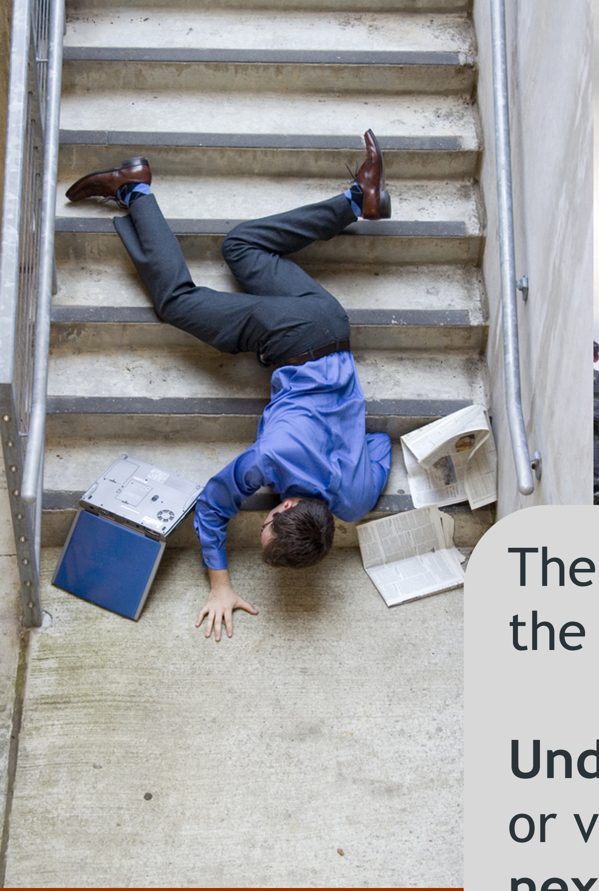


This Talk: Language comprehension in eventful contexts

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

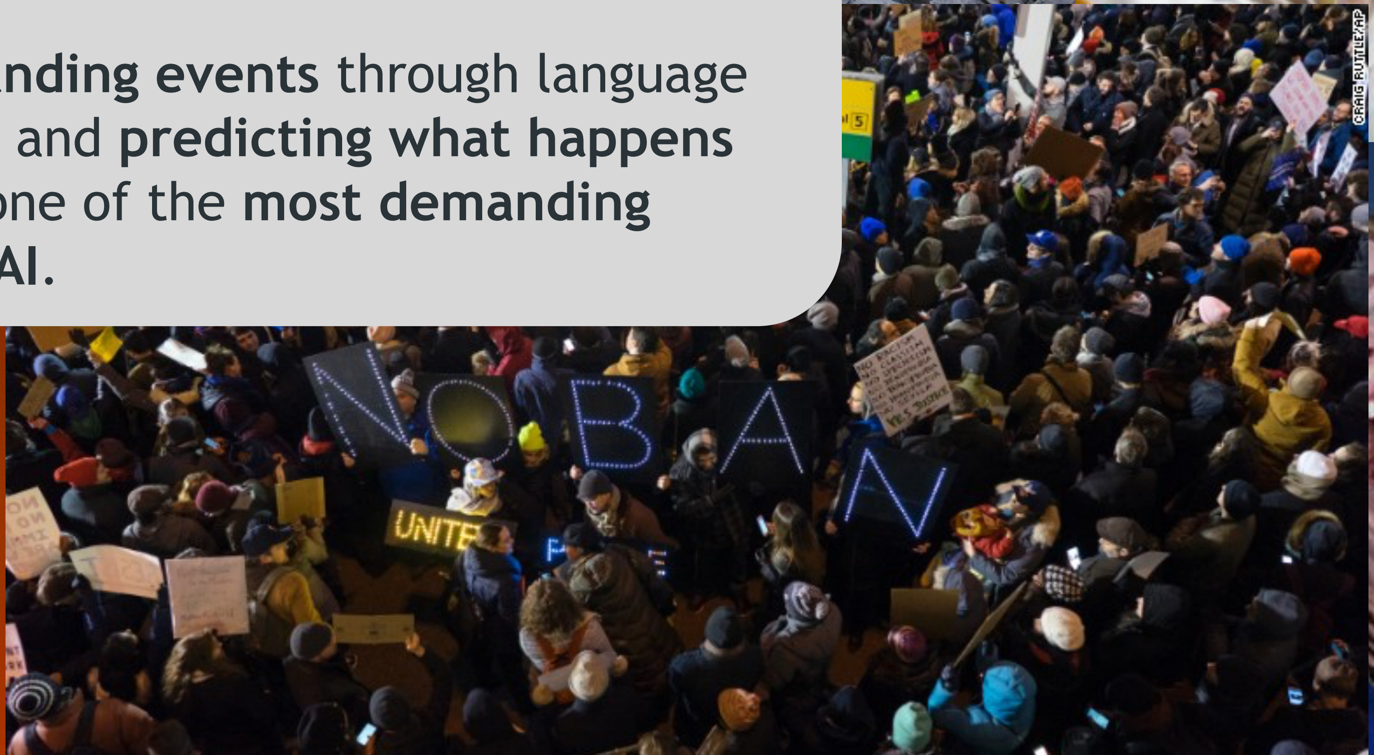
UNIVERSITY of ROCHESTER

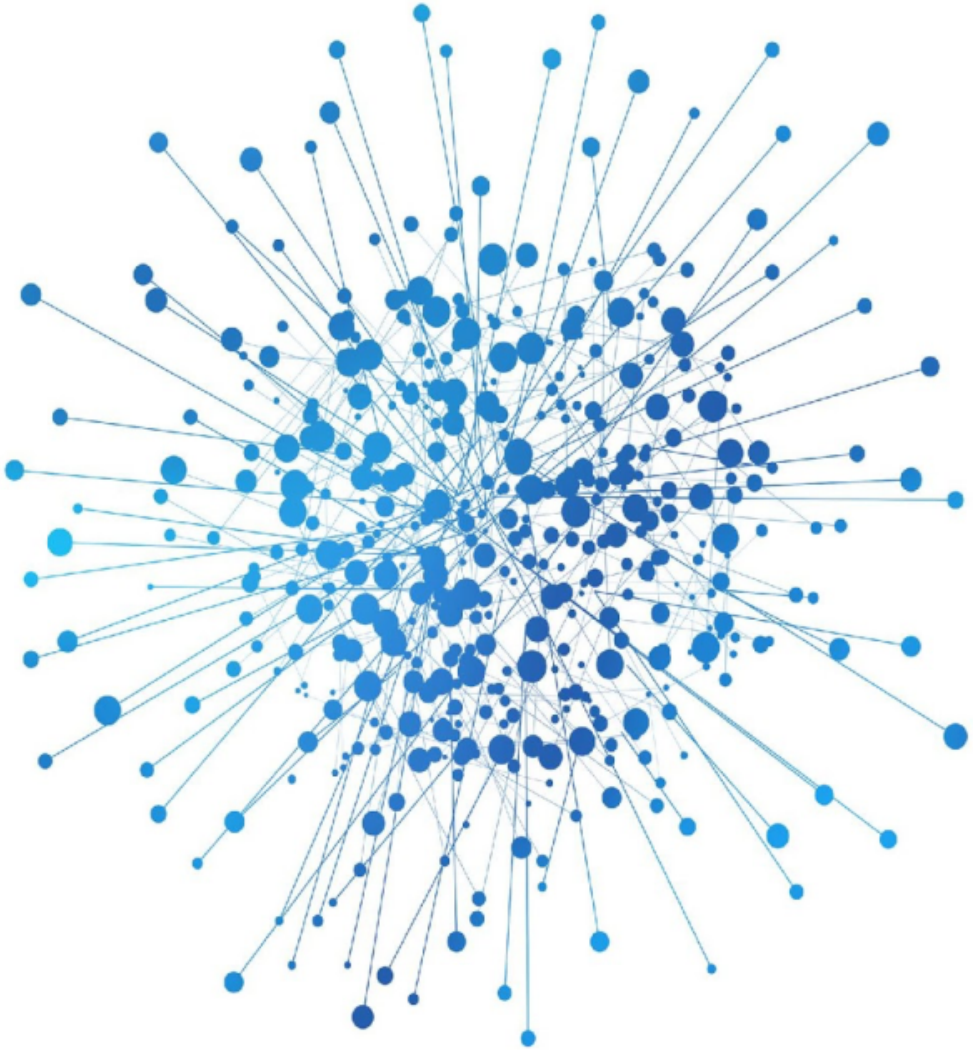
Microsoft Research



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

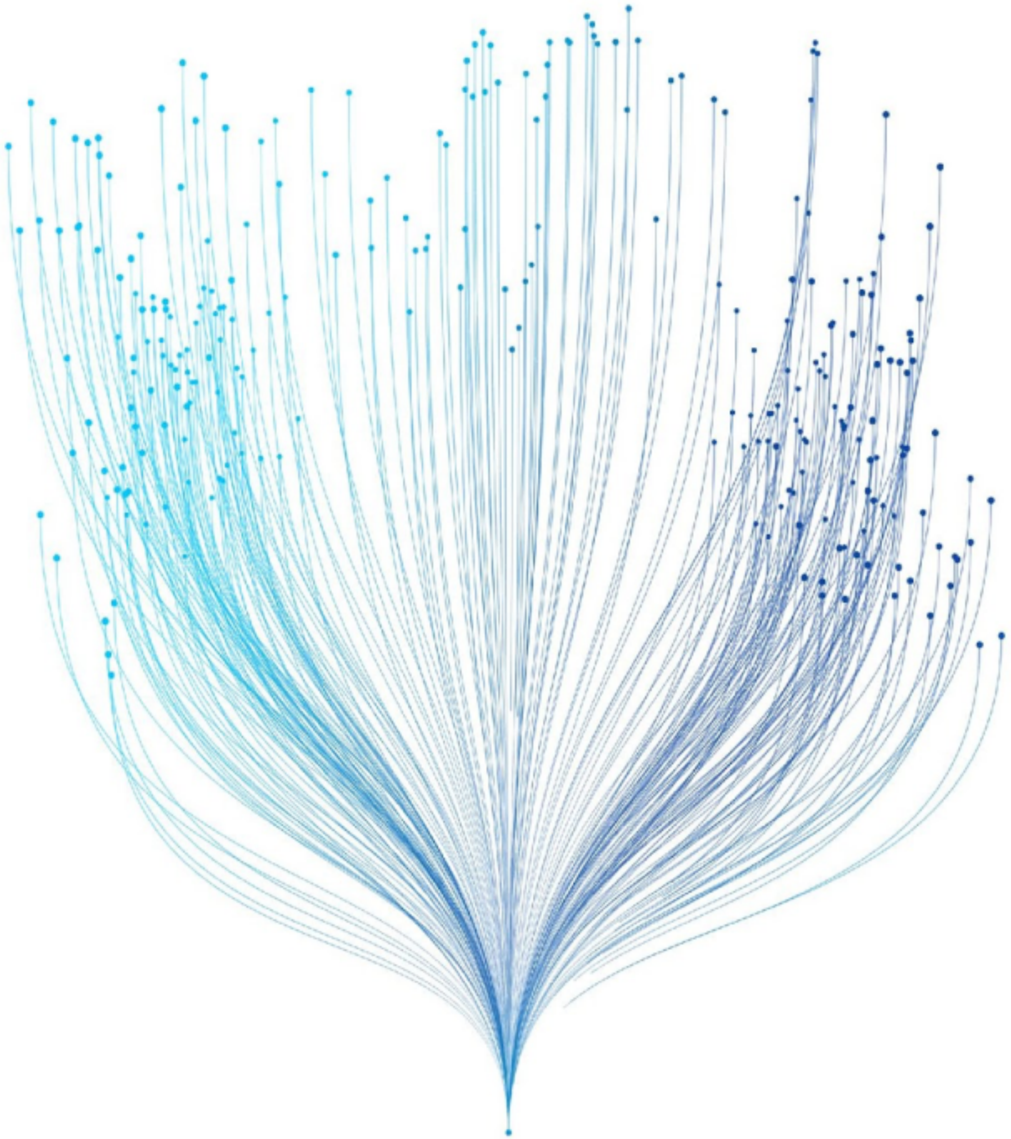
Understanding events through language or vision, and **predicting what happens next**, is one of the **most demanding areas in AI**.





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

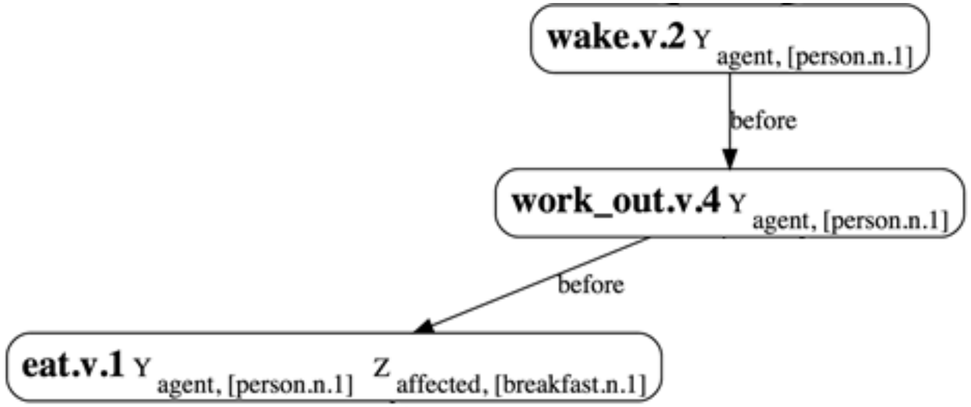
Mostafazadeh et al., NAACL 2016

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.


wake.v.2 Y agent, [person.n.1]
before
work_out.v.4 Y agent, [person.n.1]
before
eat.v.1 Y agent, [person.n.1] Z affected, [breakfast.n.1]

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

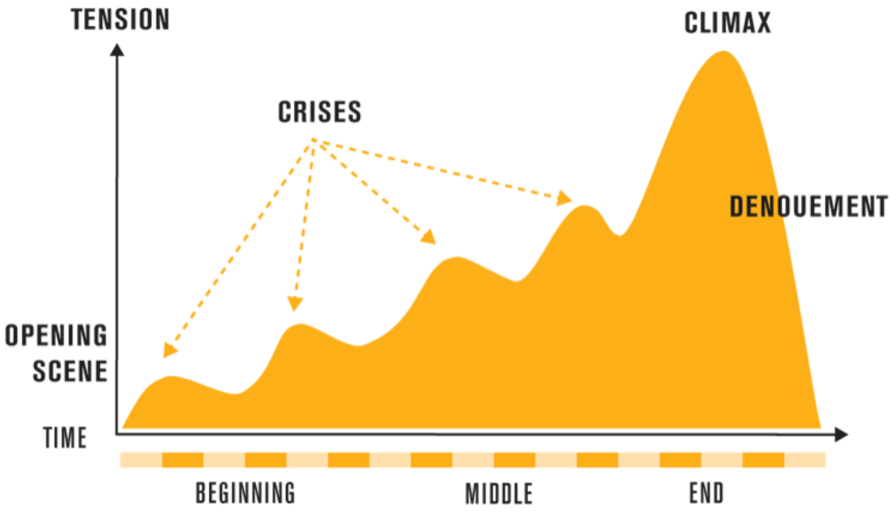
ROCStories

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

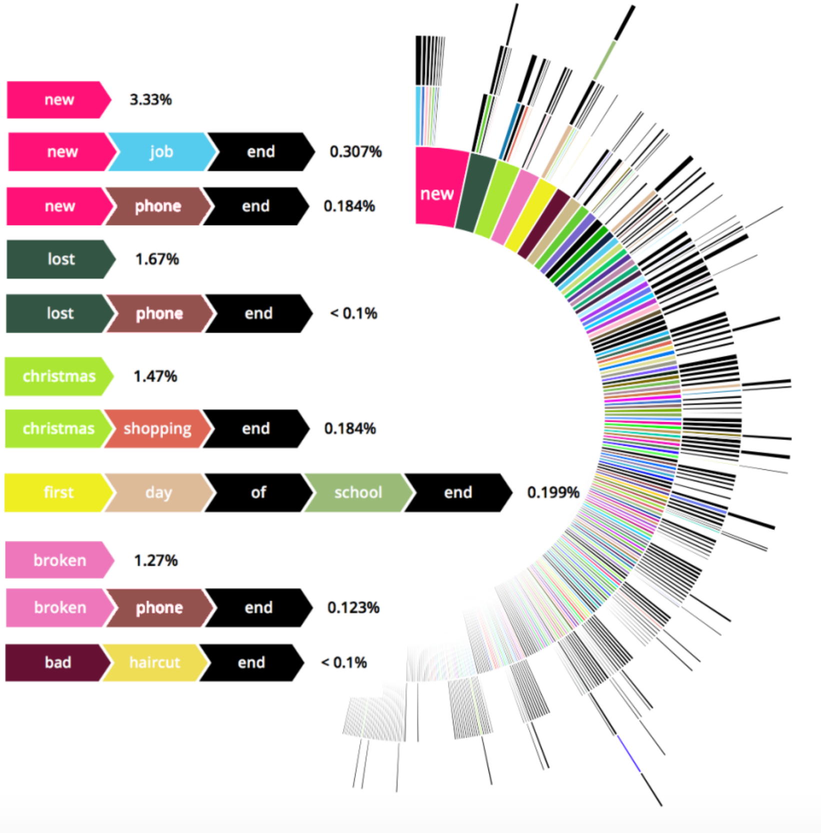
Characteristics:

- 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”

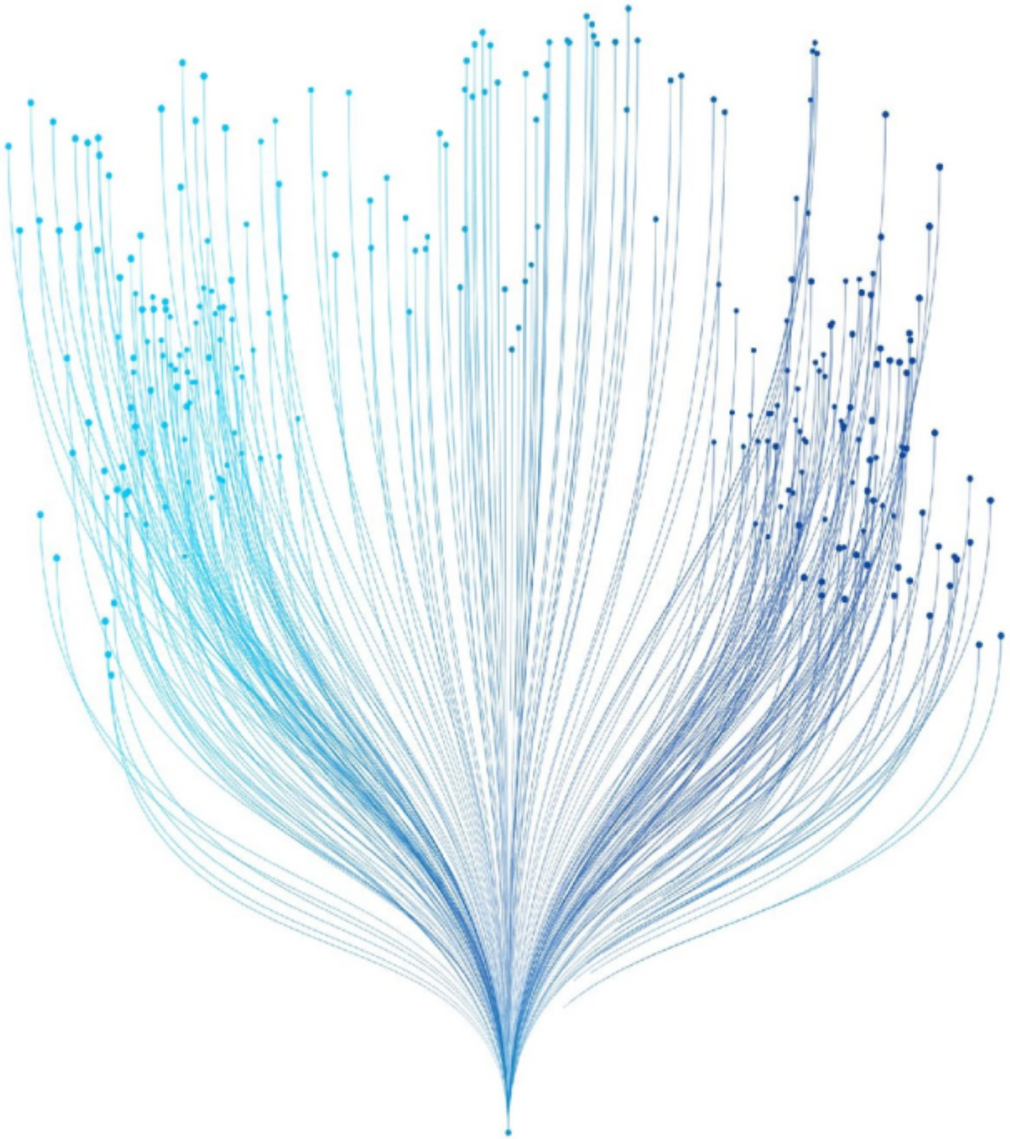
- Bill thought he was a great basketball player. He challenged Sam to a friendly game. He agreed. Sam started to practice really hard. Eventually, Sam beat Bill by 40 points.

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!

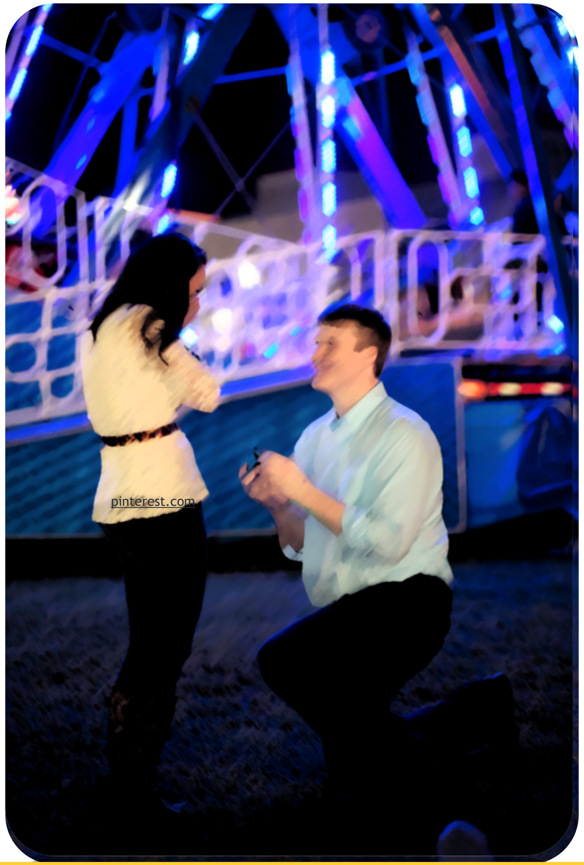
Our Idea: Story Cloze Test (SCT)

- **Goal:** Design a new evaluation schema for story understanding and narrative structure learning.

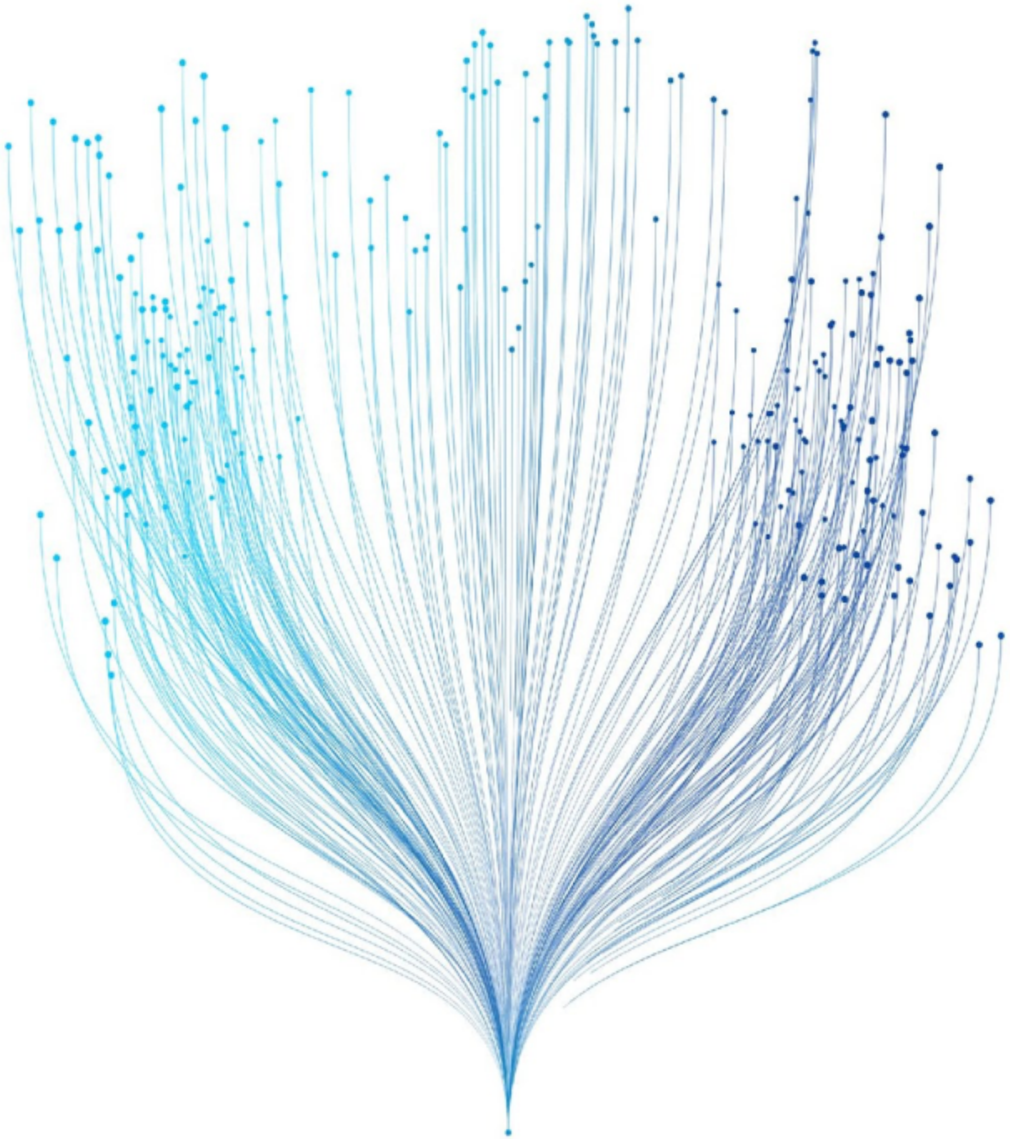
- **The Story Cloze Test:** Given a context of four sentences, predict the ending of the story.
  - Collect this evaluation dataset of by crowdsourcing

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 wiped mud off of his boot.



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



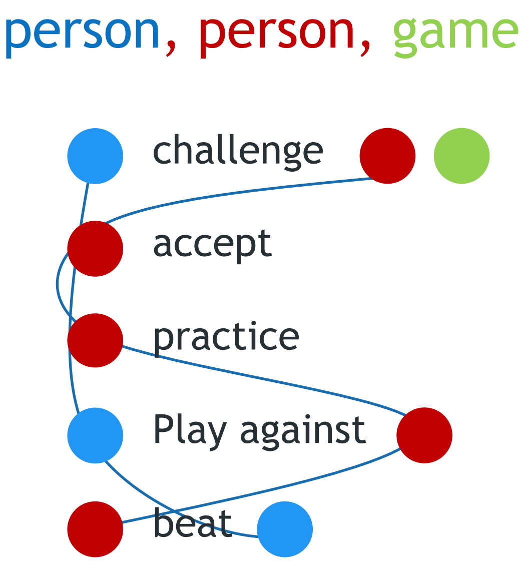
Story Cloze Models

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



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

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

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

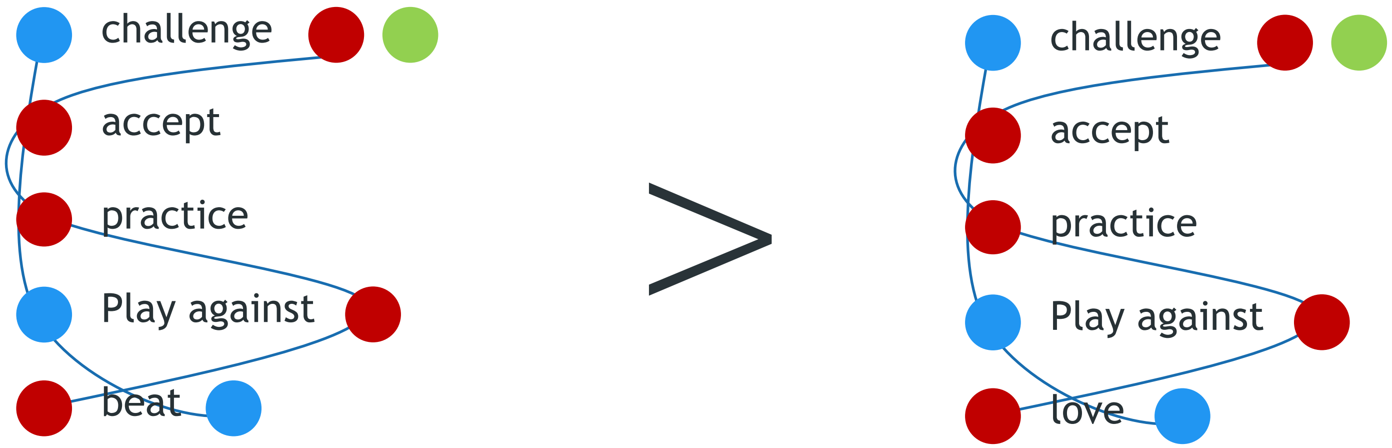
$$sim(\langle e,d\rangle,\langle v,d\rangle,a) = pmi(\langle e,d\rangle,\langle v,d\rangle) + \lambda \log C(\langle e,d\rangle,\langle v,d\rangle,a)$$

Chambers & Jurafsky, ACL 2009

Learning Typed Narrative Schemes 2/2

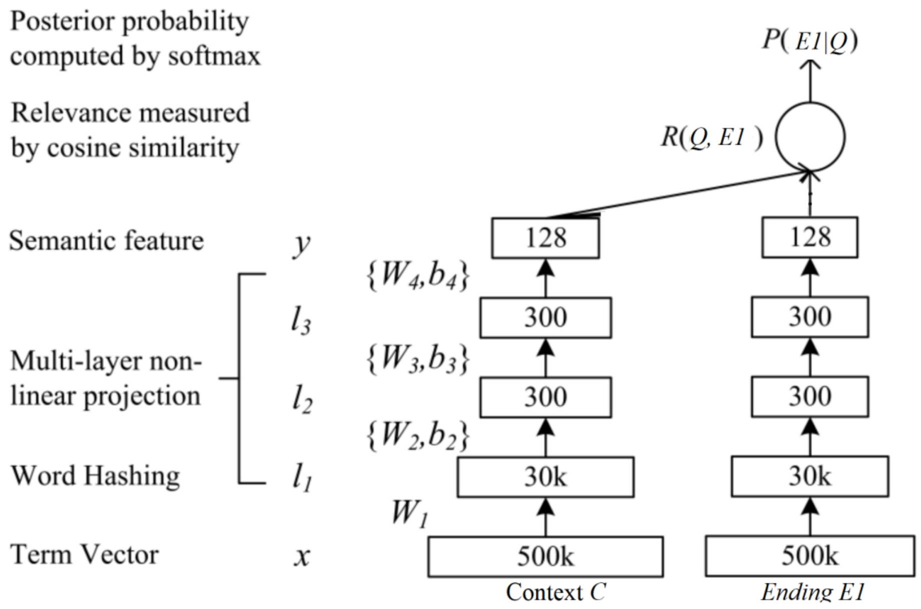
At test time

- 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.

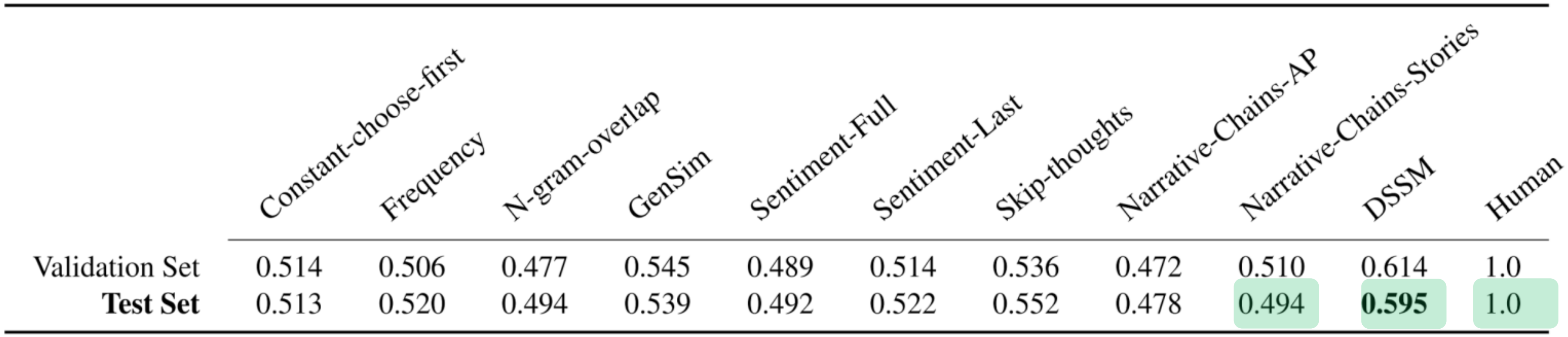


Baseline Models

- **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

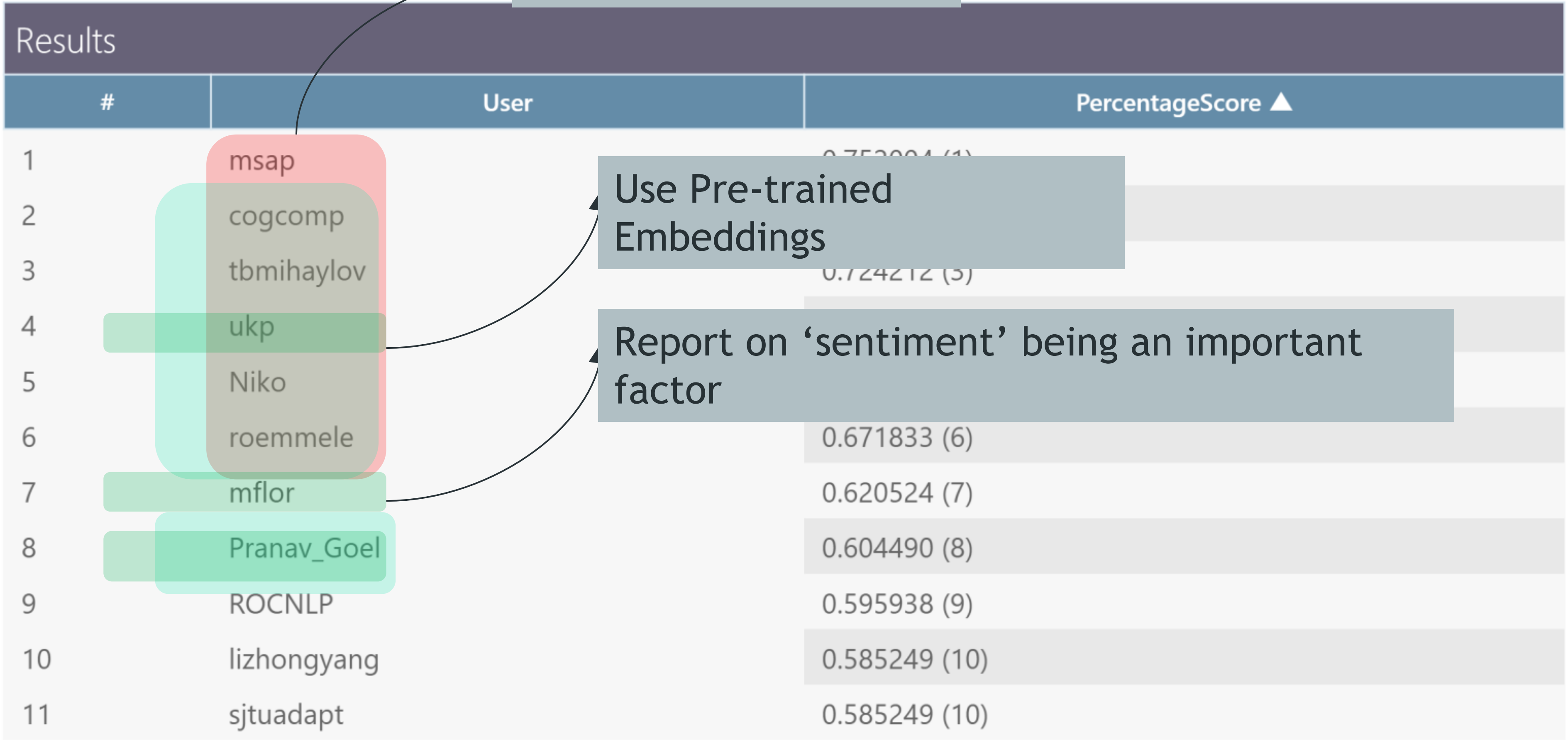
- Human performs 100%
- A challenging task with a wide enough gap (42%) from the state-of-the-art and human performance, so plenty of room for research!

- 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
    - The participants were encouraged to use any training data of their choice
- A variety of submitted approaches
  - Rule-based methods
  - Linear classifiers using different discourse phenomena
  - End-to-end neural models
  - Hybrid models

Notable Trends



Results

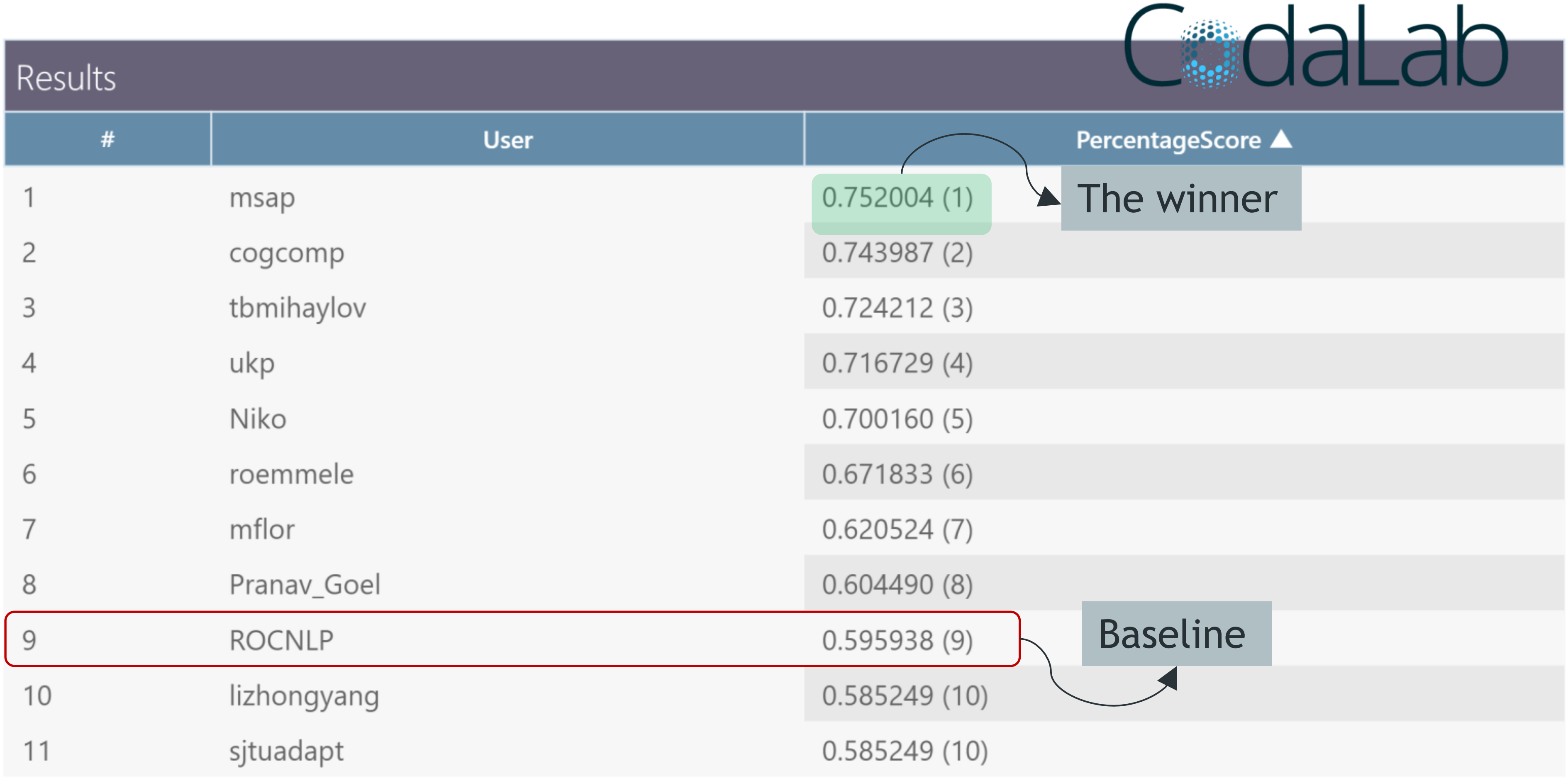
#	User	PercentageScore ▲
1	msap	0.752004 (1)
2	cogcomp	
3	tbmihaylov	0.724212 (3)
4	ukp	
5	Niko	
6	roemmele	0.671833 (6)
7	mflor	0.620524 (7)
8	Pranav_Goel	0.604490 (8)
9	ROCNLP	0.595938 (9)
10	lizhongyang	0.585249 (10)
11	sjtuadapt	0.585249 (10)

Use DNN in some way

Use Pre-trained Embeddings

Report on ‘sentiment’ being an important factor

Story Cloze Shared Task



CodaLab

Results

#	User	PercentageScore ▲
1	msap	0.752004 (1)
2	cogcomp	0.743987 (2)
3	tbmihaylov	0.724212 (3)
4	ukp	0.716729 (4)
5	Niko	0.700160 (5)
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)

The winner

Baseline

The Shared Task Winner: UW team

(Schwartz et al., 2017)

- A submission from UW team who had previously worked on determining features such as gender or age from authorship styles.
- Logistic regression classifier using the following features
  - LSTM LM probabilities trained on full stories
  - Linguistic “stylistic” features of only the ending sentences (correlated with deceptive text features)
    - Sentence length
    - Word & character n-gram

Very crucial to discover hidden data biases in various AI tasks:

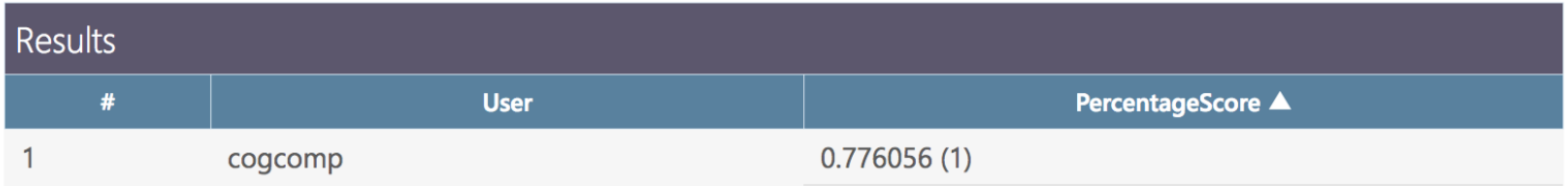
Although we were very careful with our **task design, data collection process, and establishing various baselines**, this model suggested that our **writing task** has imposed its own biases on our dataset.

Check out: The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task, Schwartz et al., CoNLL 2017

Current SOTA, UIUC team

Story Comprehension for Predicting What Happens Next

EMNLP'17



Results

#	User	PercentageScore ▲
1	cogcomp	0.776056 (1)

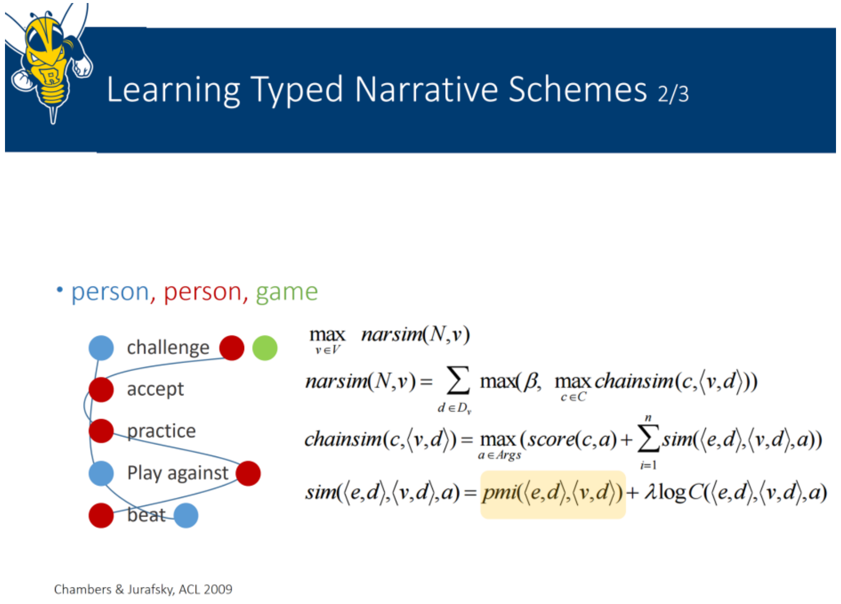
Hey, Juicero!

JUICERO
JUICERO
CHOPPED
Greens
FRESH CHOPPED PRODUCE FOR MAKING COLD-PRESSED JUICE
NET WEIGHT 10 OZ (283 g)

Beautiful Engineering

& Our Obsession with Complexity …

Our Love for Model Complexity… 1/2



Learning Typed Narrative Schemes 2/3

- person, person, game

challenge
accept
practice
Play against
beat

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

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

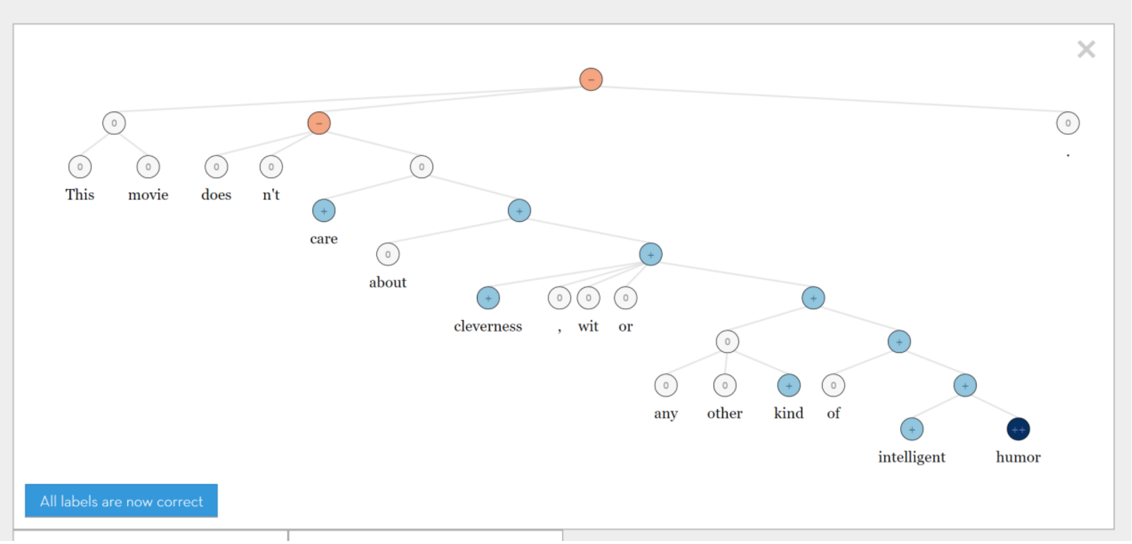
$$chainsim(c,\langle v,d \rangle) = \max_{a \in Args} (score(c,a) + \sum_{i=1}^{n} sim(\langle e,d \rangle,\langle v,d \rangle,a))$$

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

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

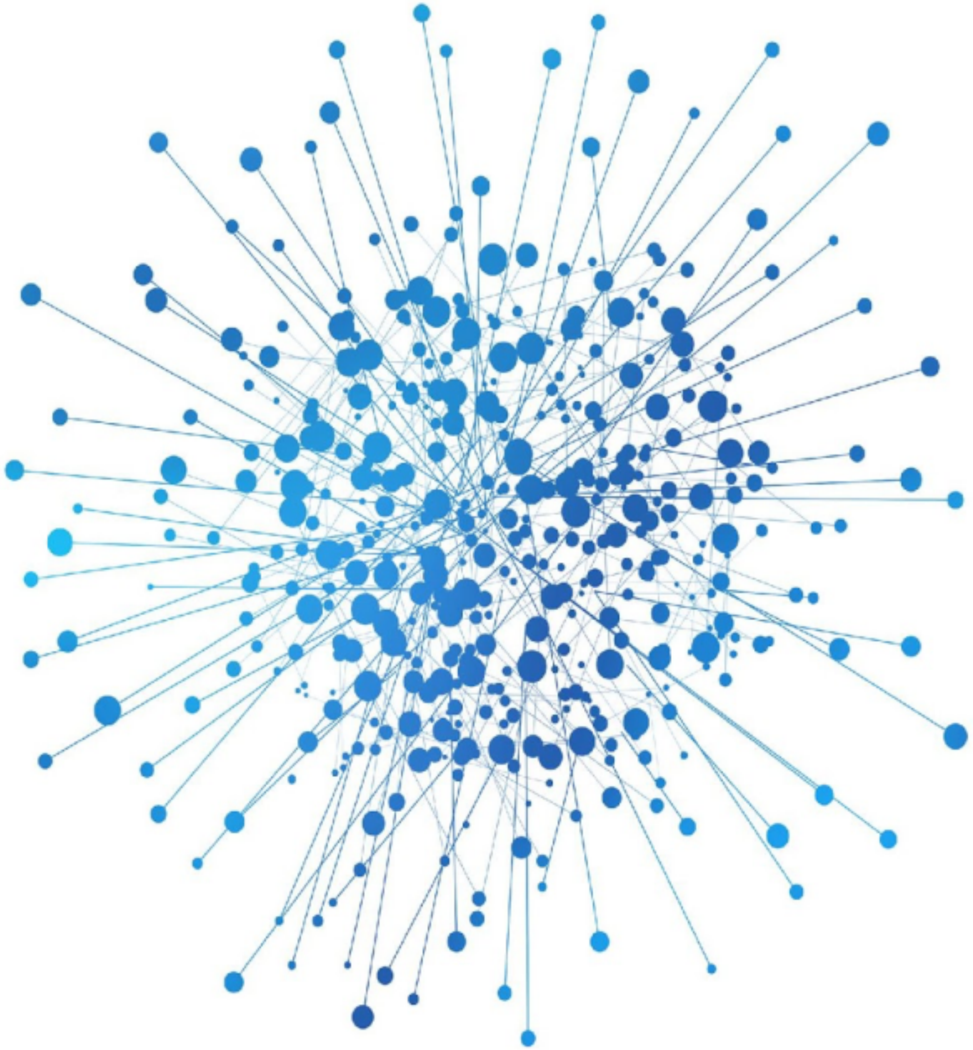


We used "Recursive Neural Network" sentiment analyzer trained on ~12,000 sentences and achieved **49.2**

- 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/



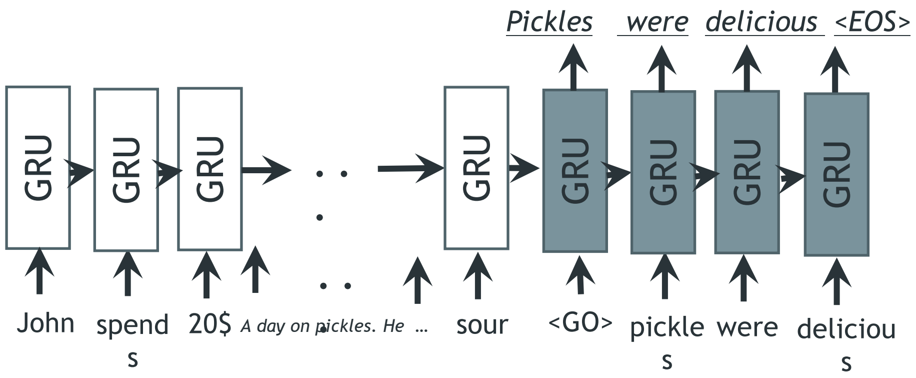
Collaborative Story Composition

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

Sequence2Sequence generation

(Sutskever et. al, 2014)

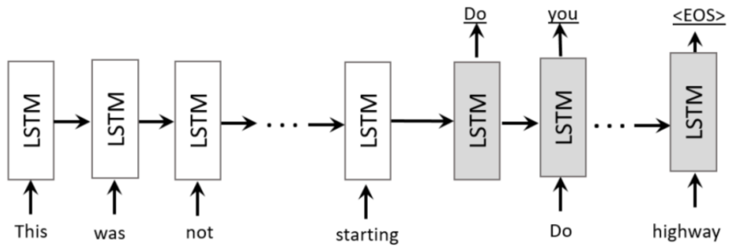
- Encoder-Decoder RNN Architecture with Attention
- 2-layers, with 512 units per layer
- Beam-search decoding, with beam-width = 25



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 .

- 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

- **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 !

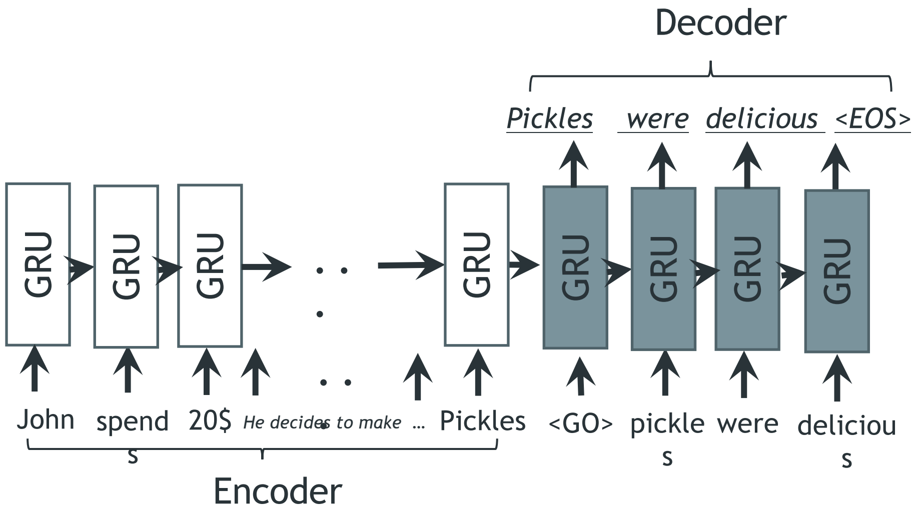
The importance of Modeling (non-trivial) Context

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>


Decoder
Pickles were delicious <EOS>
GRU
GRU
GRU
GRU
GRU
GRU
GRU
GRU
John spends 20$ He decides to make ... Pickles
<GO> pickles were delicious
Encoder

<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 phone.

- **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.

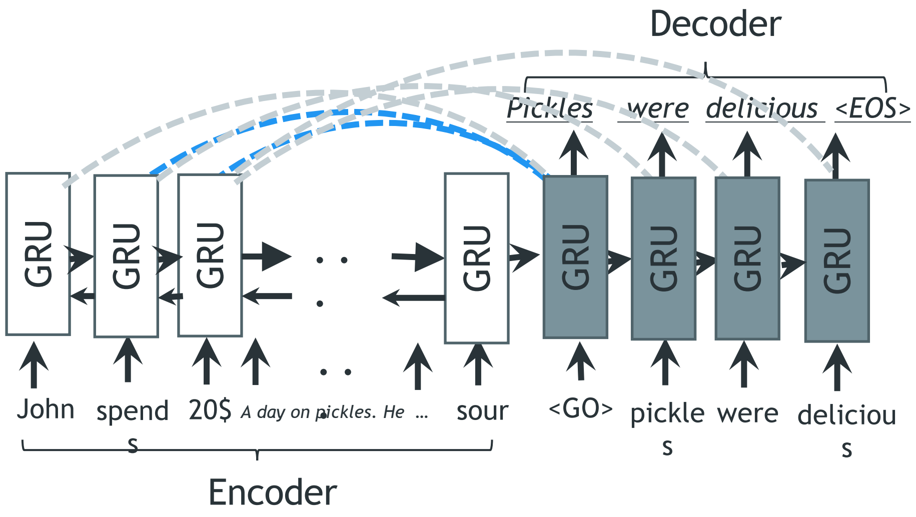
Better Context Representation

- 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?

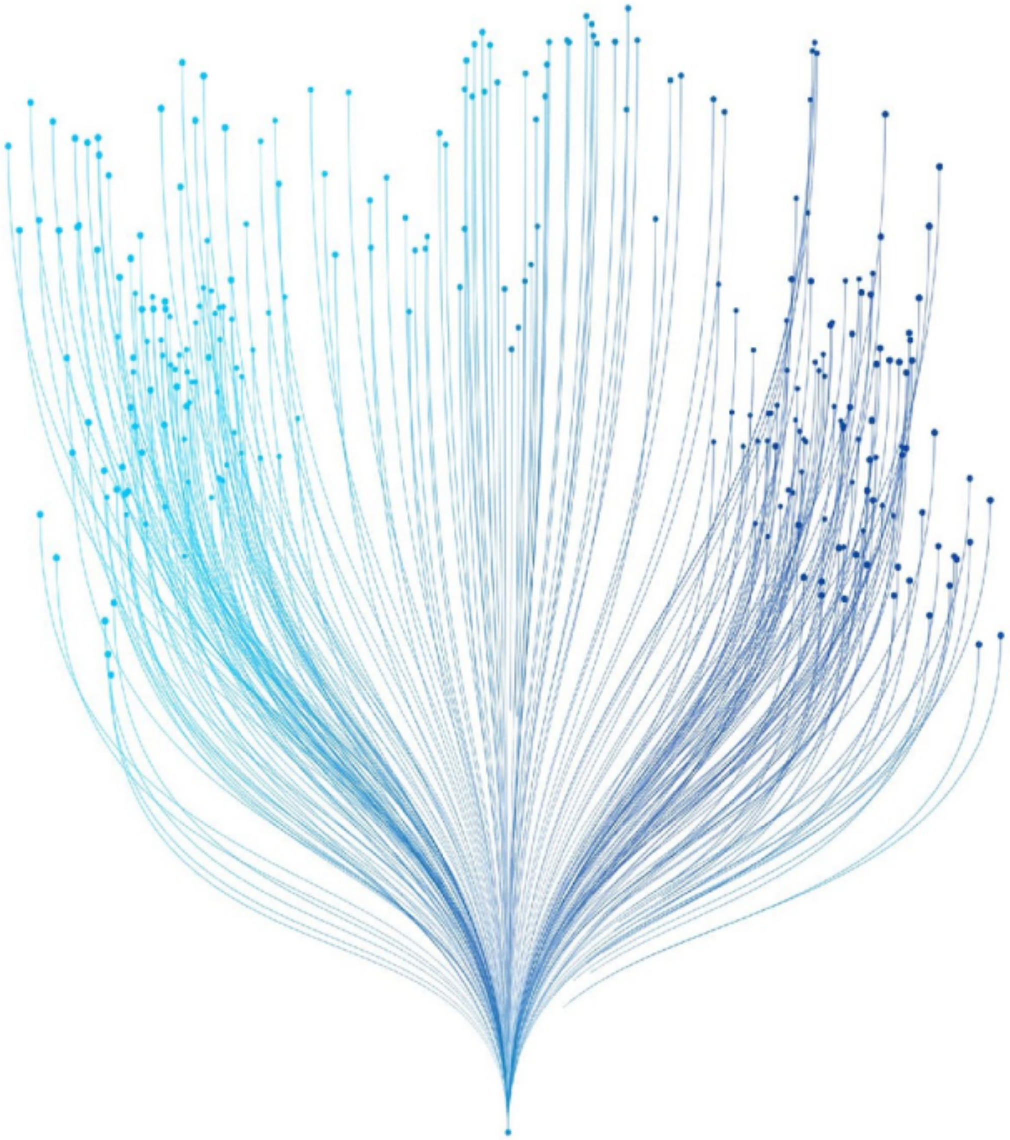
- RNNLMs are performing great on generating grammatical outputs
  - Local coherency
- **Logically-sound** generation is still very challenging
  - Generation given a trivial context (a topic, or a title) is easier than generating a **logically-sound output** given **a non-trivial long context**
- Generating Shakespeare-like text or poetry is not as challenging
  - Often, irrelevant content can be 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?

Is the motorcyclist alive?

What happened?

Was anyone injured in the crash?

Is anyone injured?

Is the motorcyclist all right?

What happened?

Is the motorcyclist OK?

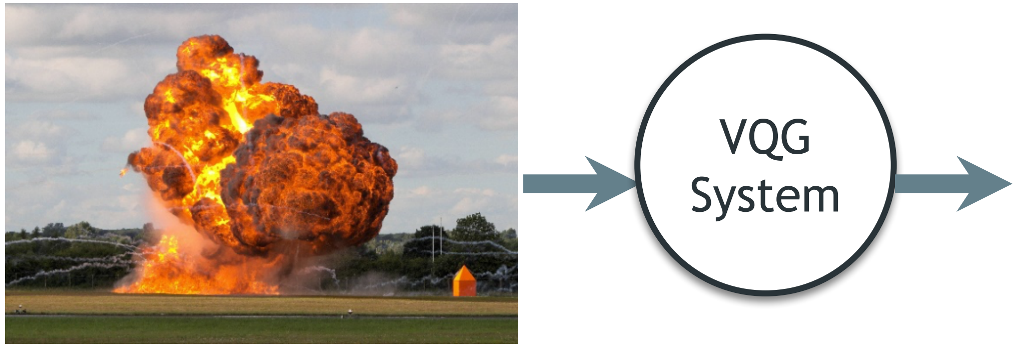
What caused this accident?

Was anyone injured in the crash?

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

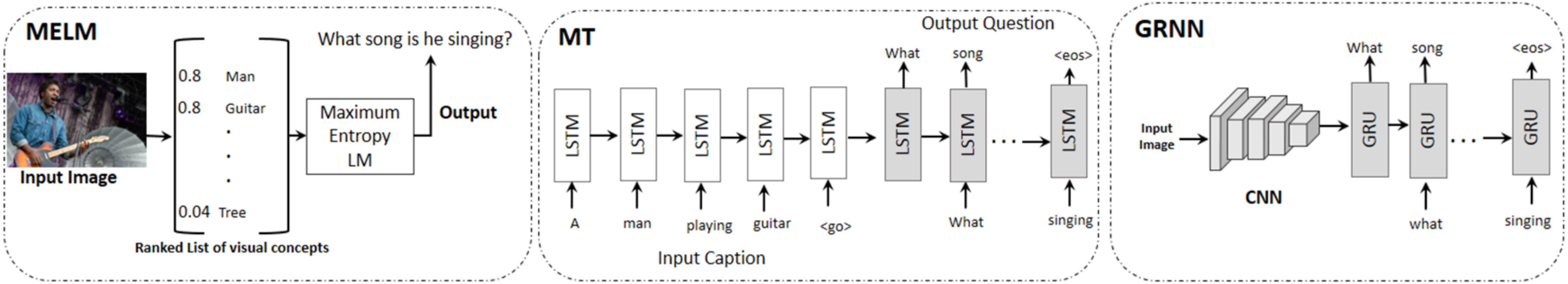
Models



VQG System

What is being burned here?

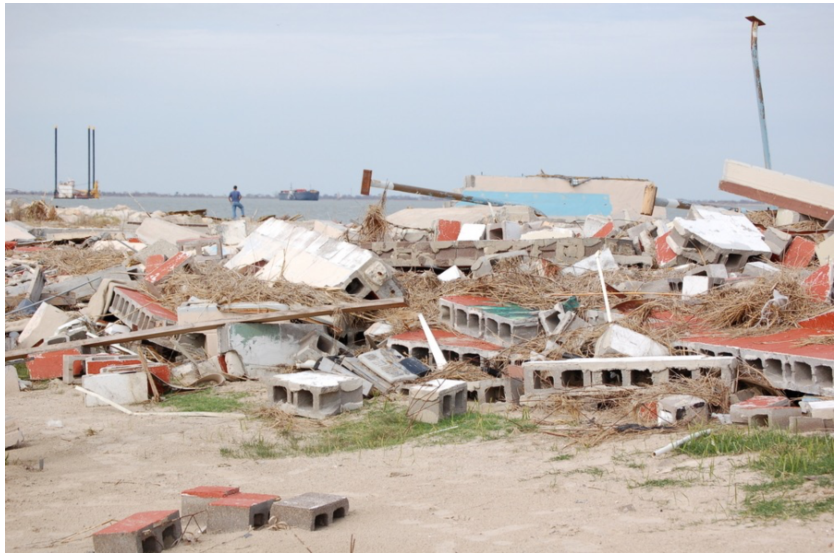
Generation Models & Results



	$Human_{consensus}$	$Human_{random}$	$GRNN_X$	$GRNN_{all}$	$1\text{-}NN_{bleu-X}$	$1\text{-}NN_{gensim-X}$	$K\text{-}NN{+}min_{bleu-X}$	$K\text{-}NN{+}min_{gensim-X}$	$1\text{-}NN_{bleu-all}$	$1\text{-}NN_{gensim-all}$	$K\text{-}NN{+}min_{bleu-all}$	$K\text{-}NN{+}min_{gensim-all}$
				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

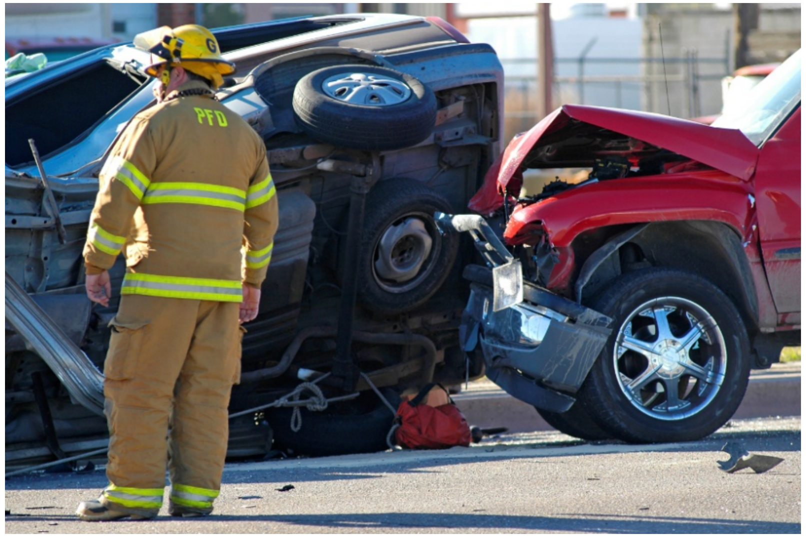
	METEOR	BLEU	ΔBLEU
r	**0.916** (4.8e-27)	0.915 (4.6e-27)	0.915 (5.8e-27)
ρ	0.628 (1.5e-08)	0.67 (7.0e-10)	**0.702** (5.0e-11)
τ	0.476 (1.6e-08)	0.51 (7.9e-10)	**0.557** (3.5e-11)

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

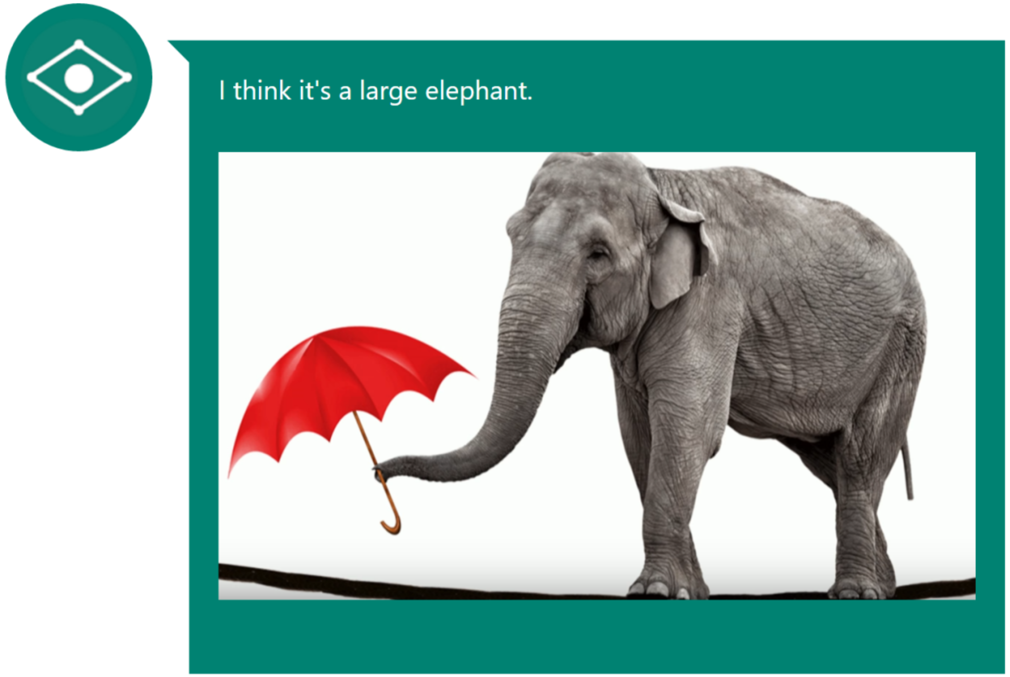


- Did the drivers of this accident live through it?
- **GRNN**: How did the car crash?
- **KNN**: Was anybody hurt in this accident?
- **Caption Bot**: A man standing next to a motorcycle.

Image Captioning

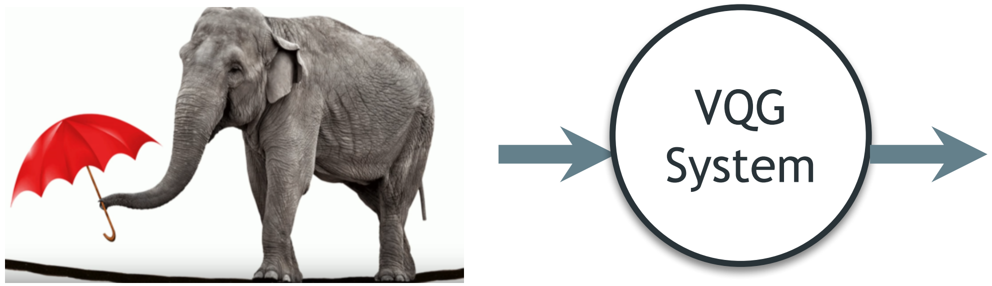
Out of the scope of the training data

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

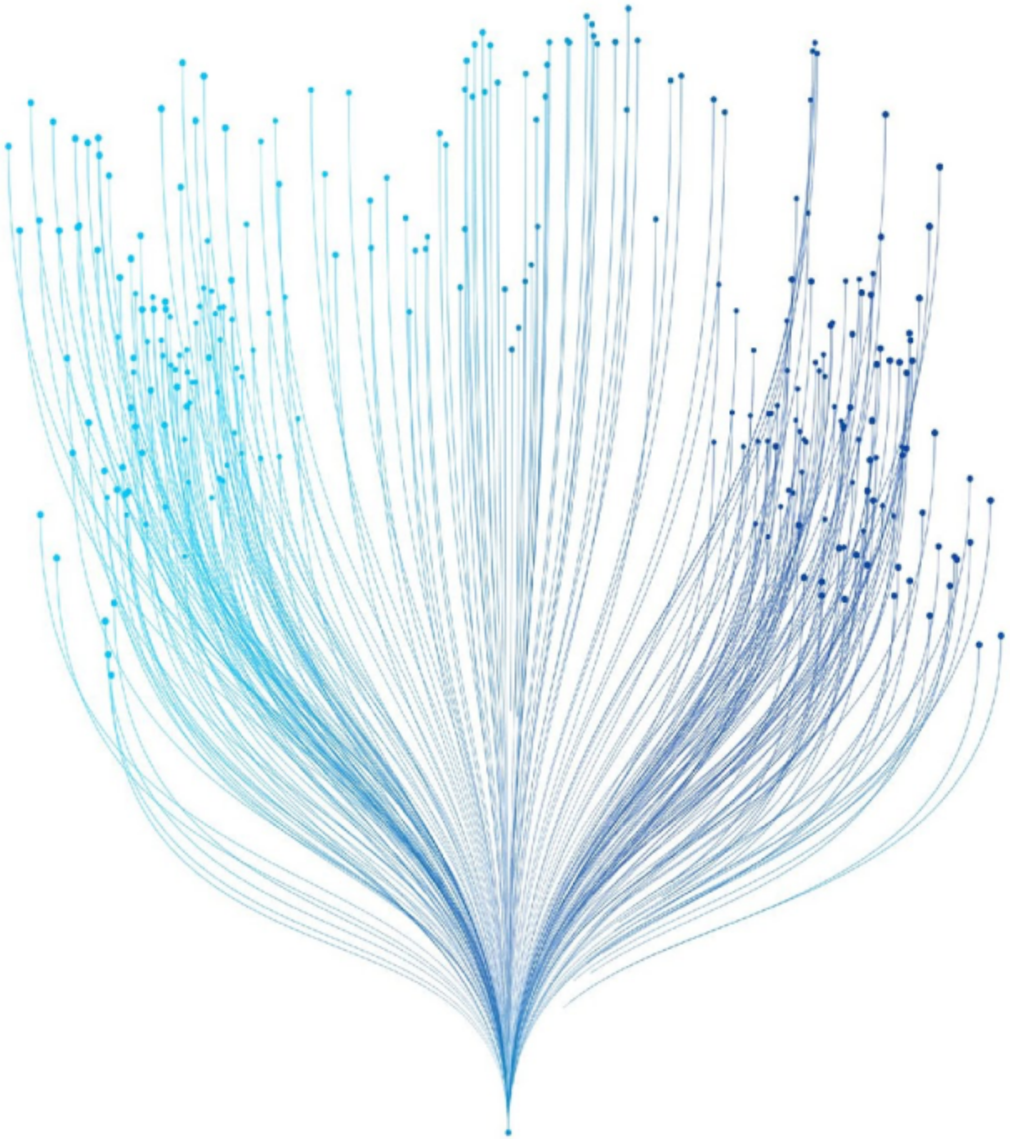

I think it's a large elephant.

Visual Question Generation

Out of the scope of the training data


VQG System

What kind of animal is this?

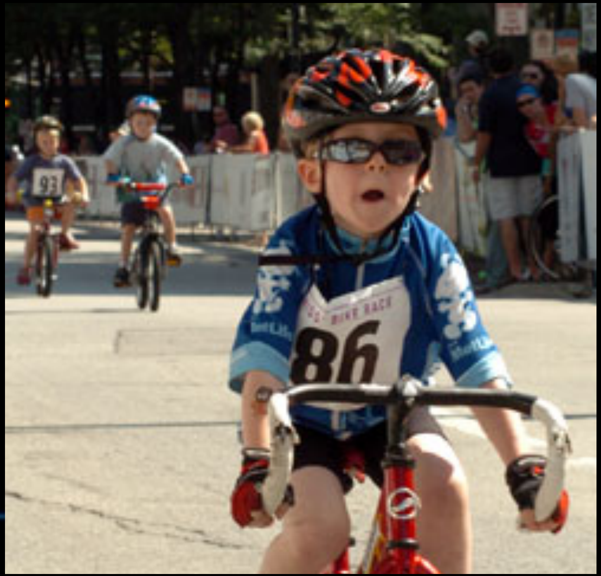


3. Modeling Visual & Textual Context

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

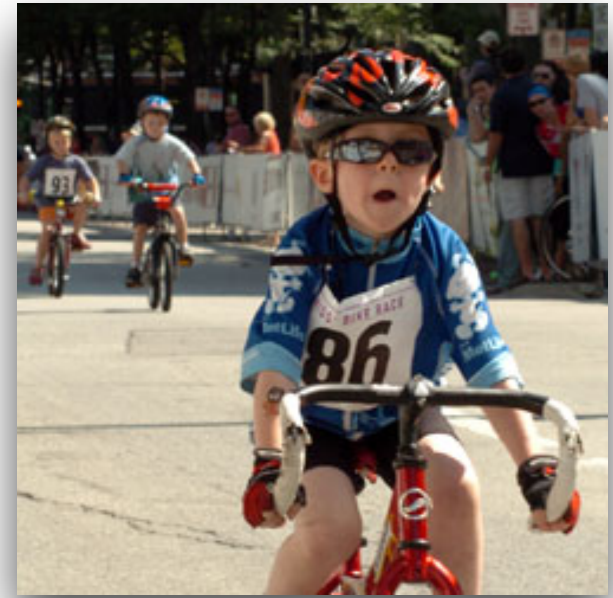
Mostafazadeh et al., IJCNLP 2017

😎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

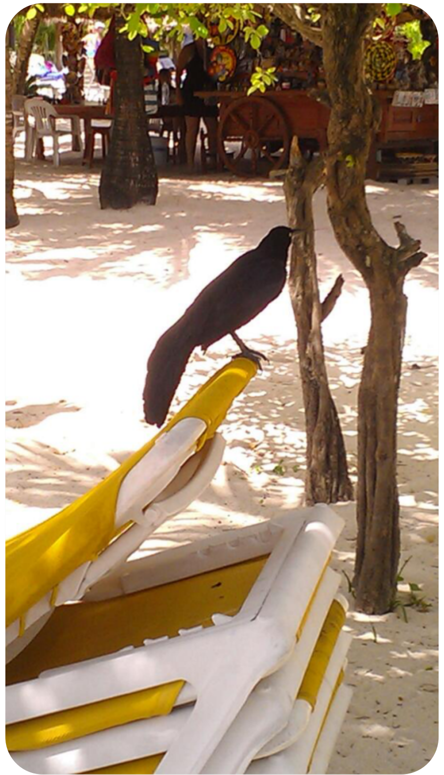
Proactively drive the conversation forward by asking “reasonable” questions!

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



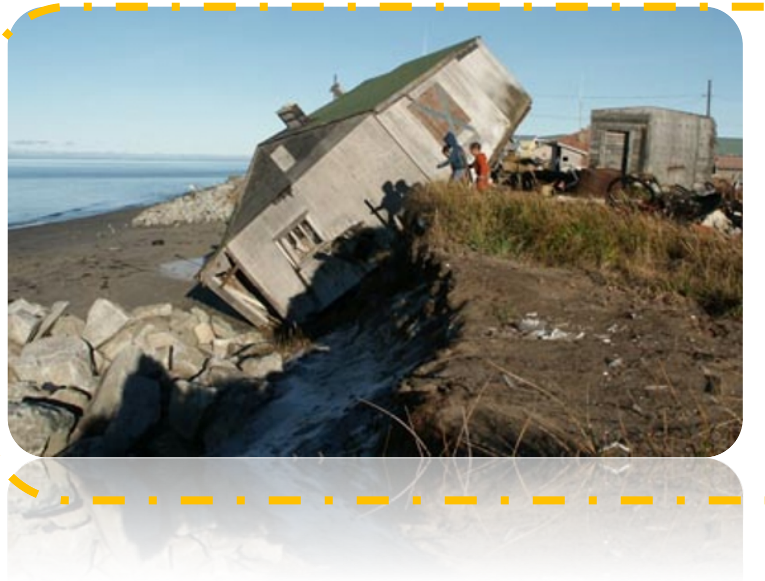
I found a cawaii bird.

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

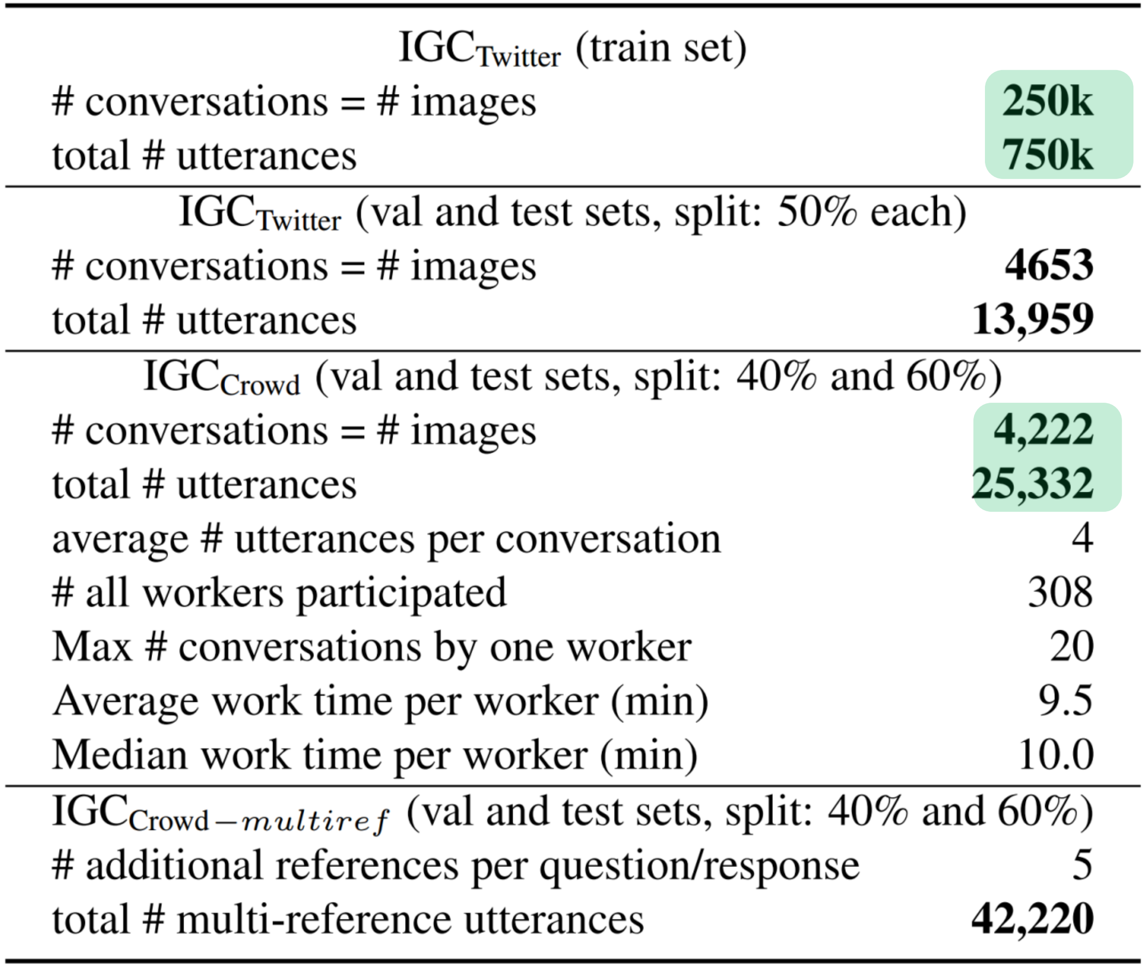


A terrible storm destroyed my house!

OH NO, what are you going to do?

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

Dataset Statistics

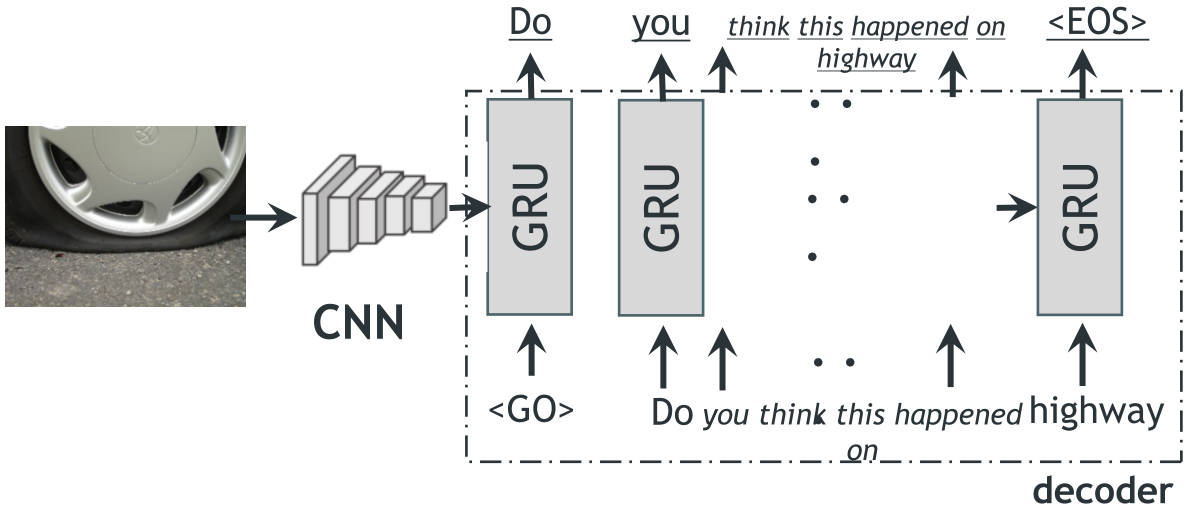


IGC$_{\text{Twitter}}$ (train set)	
# conversations = # images	**250k**
total # utterances	**750k**
IGC$_{\text{Twitter}}$ (val and test sets, split: 50% each)	
# conversations = # images	**4653**
total # utterances	**13,959**
IGC$_{\text{Crowd}}$ (val and test sets, split: 40% and 60%)	
# conversations = # images	**4,222**
total # utterances	**25,332**
average # utterances per conversation	4
# all workers participated	308
Max # conversations by one worker	20
Average work time per worker (min)	9.5
Median work time per worker (min)	10.0
IGC$_{\text{Crowd}-multiref}$ (val and test sets, split: 40% and 60%)	
# additional references per question/response	5
total # multi-reference utterances	**42,220**

Dataset available for download: https://www.microsoft.com/en-us/download/details.aspx?id=55324

Models

Visual Context Sensitive



Do you think this happened on highway <EOS>

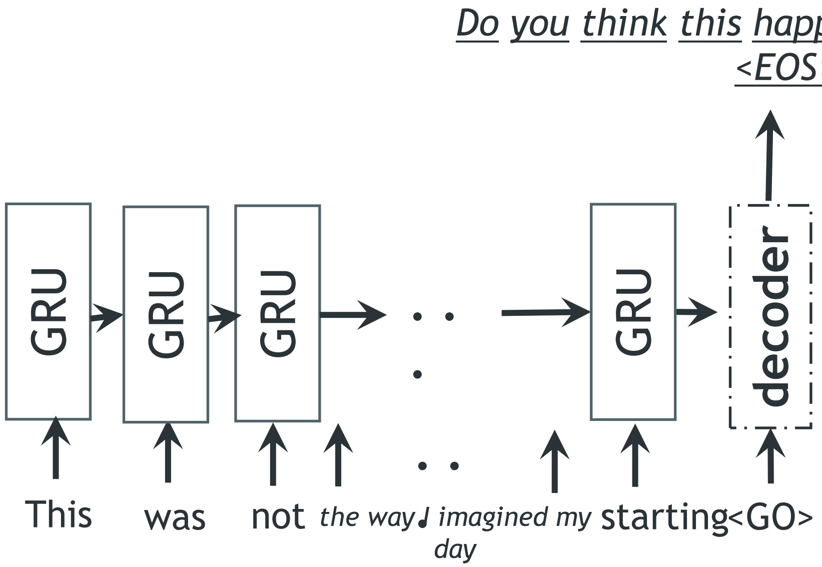
GRU GRU ... GRU

CNN

<GO> Do you think this happened on highway

decoder

Textual Context Sensitive



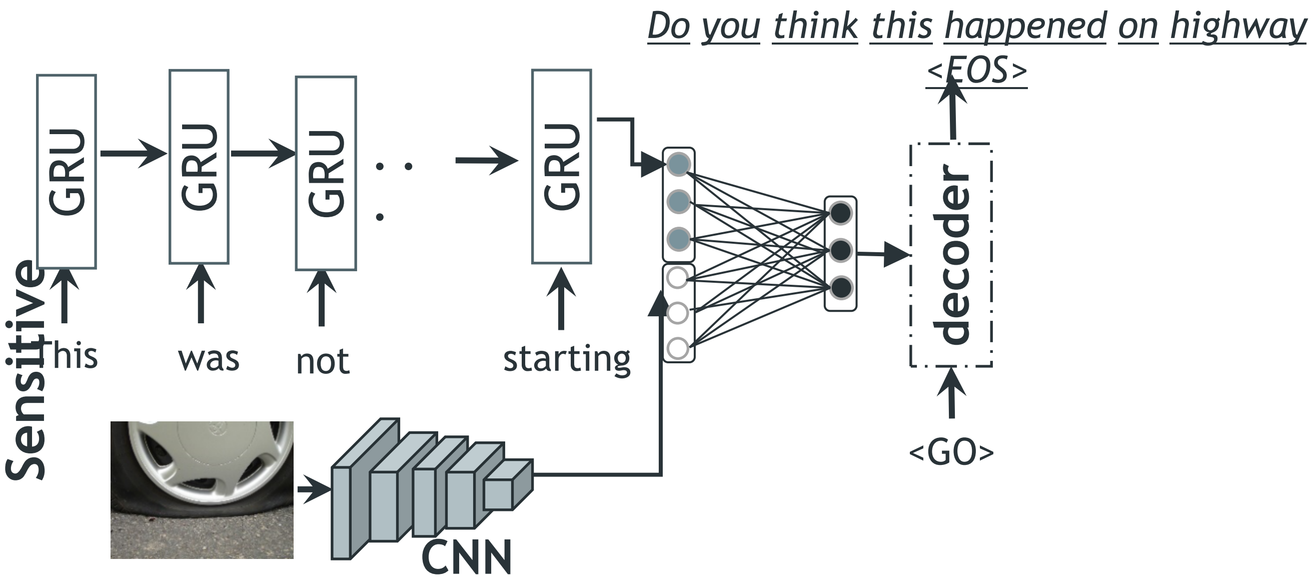
Do you think this happened on highway <EOS>

GRU GRU GRU ... GRU decoder

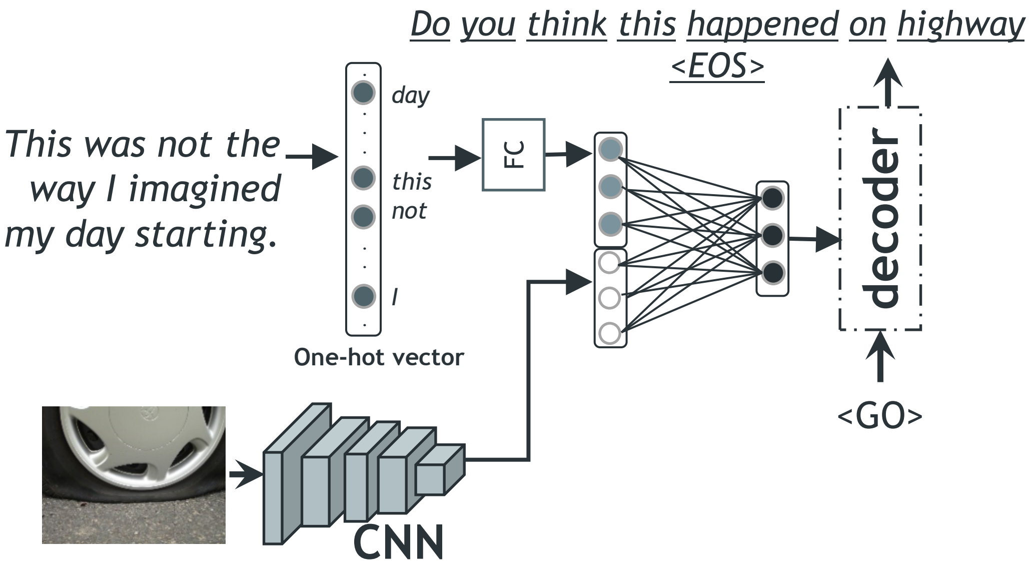
This was not the way I imagined my day starting <GO>

Models

Visual & Textual (RNN) Context Sensitive



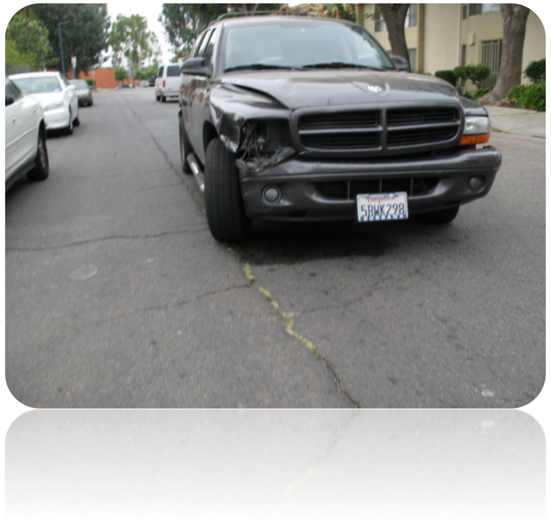
Visual & Textual (BOW) Context Sensitive



Do you think this happened on highway <EOS>
This was not the way I imagined my day starting.
day
this
not
I
One-hot vector
FC
decoder
<GO>
CNN

Image-Grounded Conversations

Question Generation



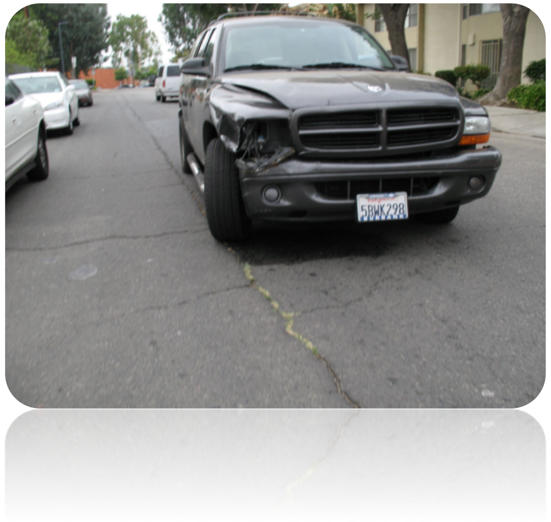
😎 I got in a car wreck today!

What happened? 🤖

😎 Well, drunk driving!

Image-Grounded Conversations

Response Generation



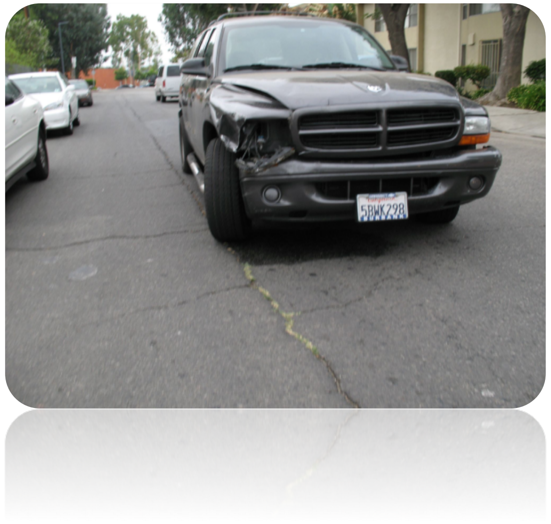
I got in a car wreck today!

Did you get hurt?

Nah, I’m home now!

Image-Grounded Conversations

Response Generation



😎 I got in a car wreck today!

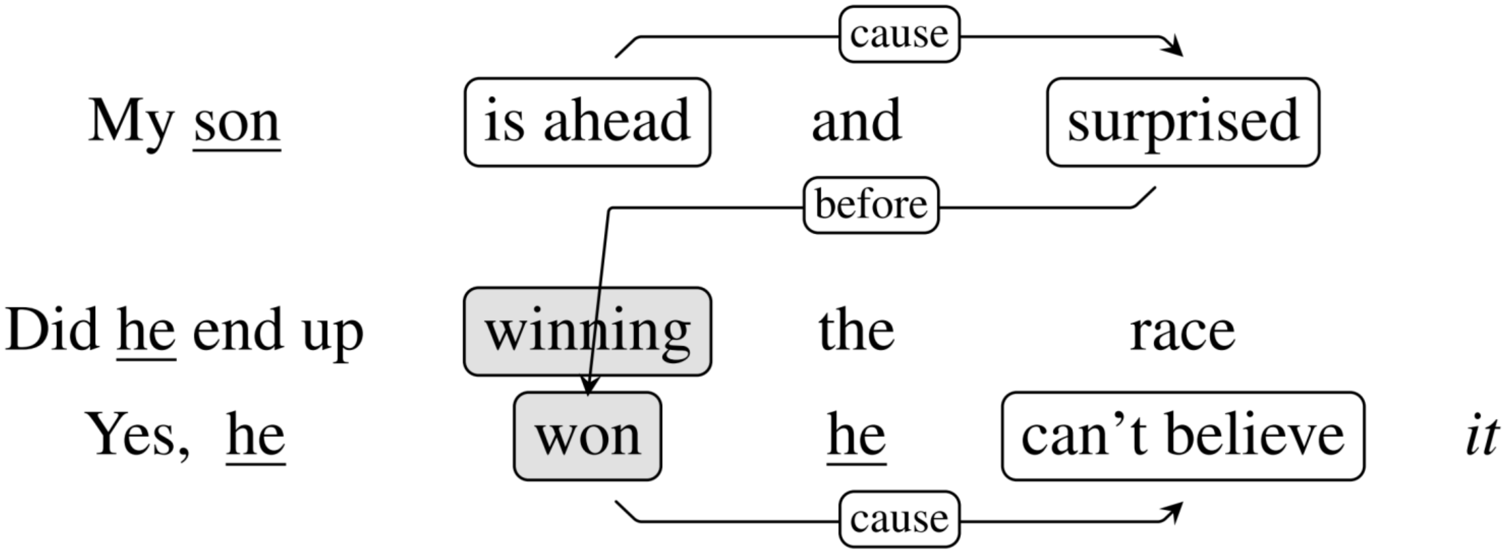
Did you get hurt? 😎

🤖 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



cause

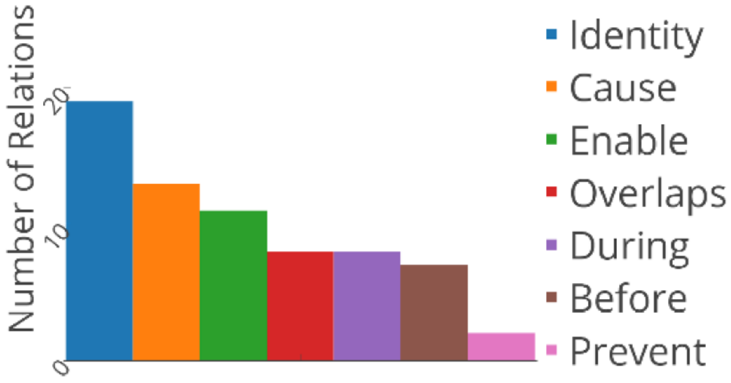
My son is ahead and surprised

before

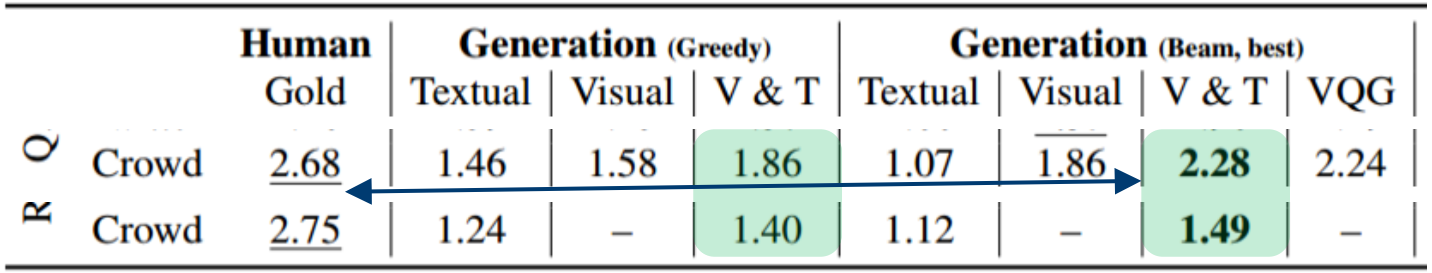
Did he end up winning the race

Yes, he won he can't believe *it*

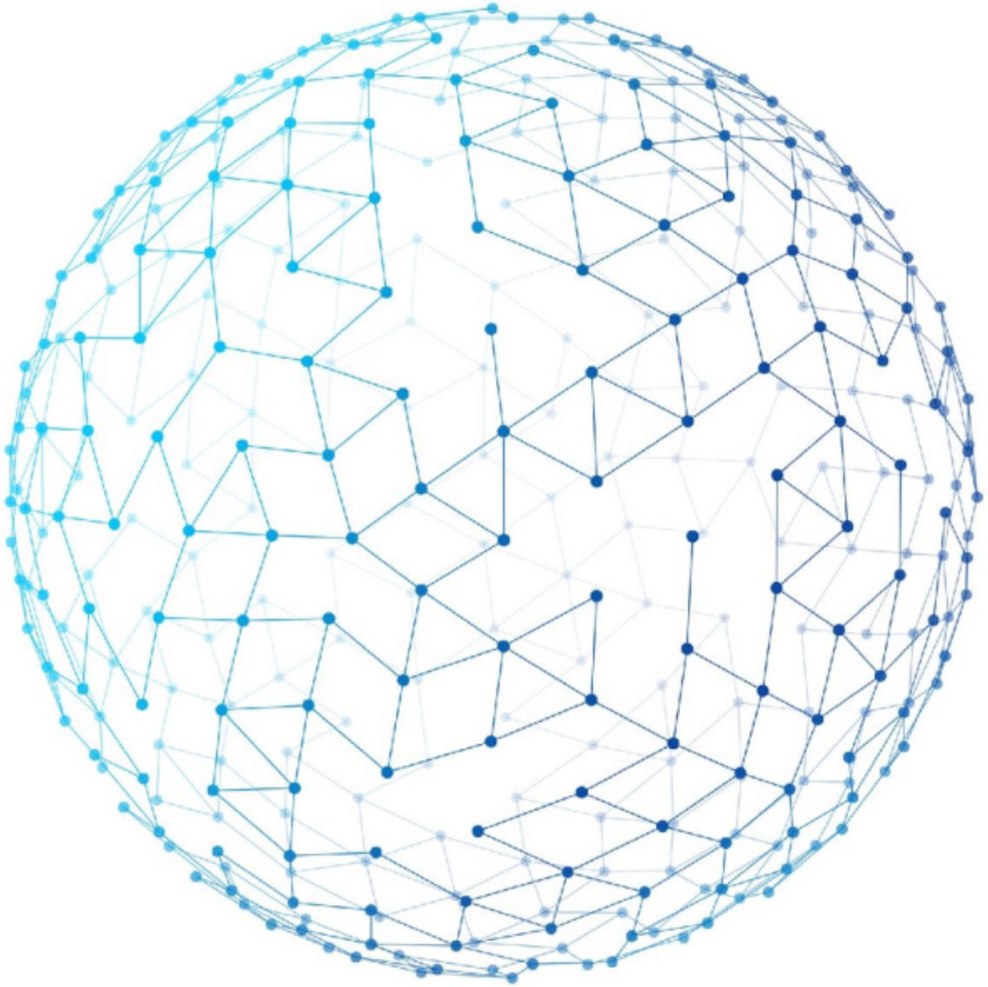
cause



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

Did the drivers of this accident live through it?

GRNN: How did the car crash?

KNN: Was anybody hurt in this accident?

Visual Question Generation

Temporal

Reading Comprehension

Event

Time

Causality

Reasoning

Knowledge Acquisition

Language Understanding

Common-sense

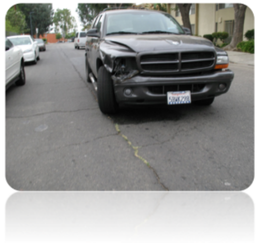
Natural Language Inference

Question Answering

Narrative Structure Learning

Story Generation

Story Understanding



I got in a car wreck today!

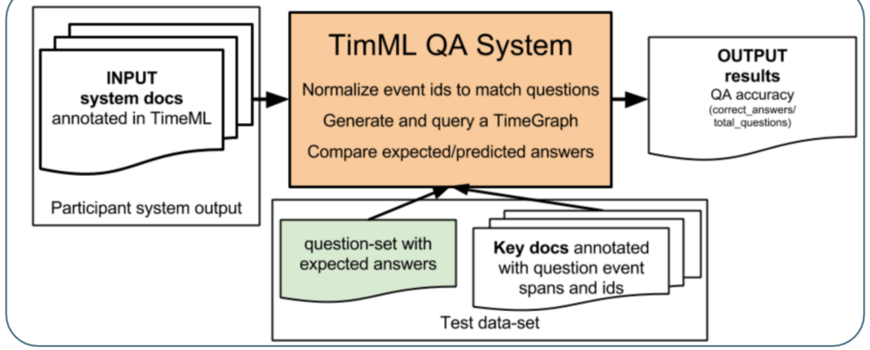
Did you get hurt?

Nah, I'm home now!

Image-Grounded Conversations

The family got together for a cookout. They had a lot of delicious food. The dog was happy to be there. They had a great time on the beach. They even had a swim in the water.

Visual Storytelling



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.

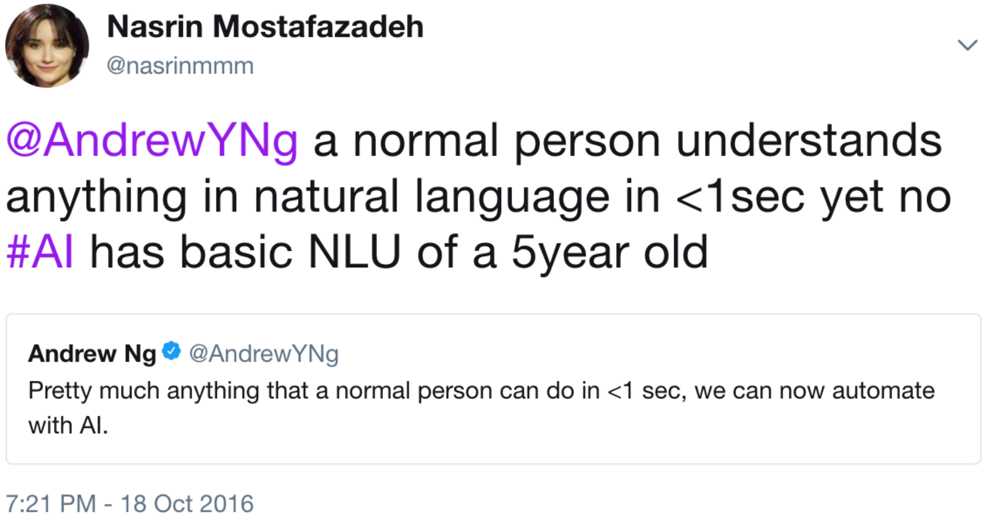
Story Cloze Test & Story Generation

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" - Yan LeCun & Gary Marcus, Nurture vs Nature in AI Debate
    - The AI community has to move towards **reasoning**



Nasrin Mostafazadeh
@nasrinmmm

@AndrewYNg a normal person understands anything in natural language in <1sec yet no #AI has basic NLU of a 5year old

Andrew Ng @AndrewYNg
Pretty much anything that a normal person can do in <1 sec, we can now automate with AI.

7:21 PM - 18 Oct 2016

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**

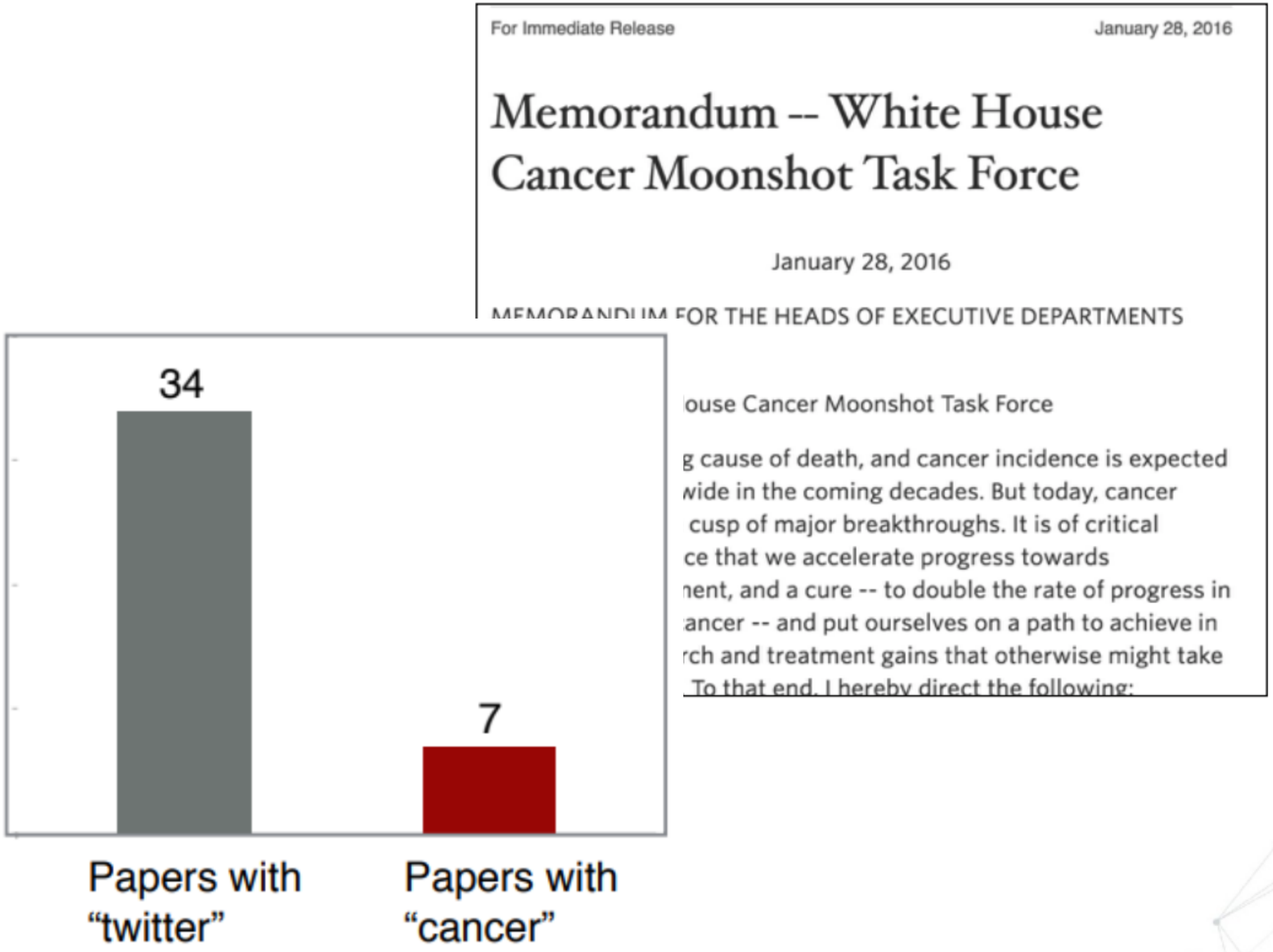
  ii. Ground predictions in a more **complex** and **realistic contexts**

  - Reasonable benchmarks with strong baselines
  - Contentful contexts are often **event-centric**

Who is (not) Fighting Cancer?

Regina Barzilay, NAACL'16 Keynote

- Bag-of-words View:
  - #Genetic 6
  - #Clinical 5
  - #Immunotherapy 4
  - #Computer Science 0
  - #Machine Learning 0
  - #Data-driven 0
  - #Big Data 0
  - **#NLP 0**



For Immediate Release January 28, 2016

Memorandum -- White House Cancer Moonshot Task Force

January 28, 2016

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS

ouse Cancer Moonshot Task Force

g cause of death, and cancer incidence is expected
wide in the coming decades. But today, cancer
cusp of major breakthroughs. It is of critical
ce that we accelerate progress towards
nent, and a cure -- to double the rate of progress in
ancer -- and put ourselves on a path to achieve in
rch and treatment gains that otherwise might take
To that end, I hereby direct the following:

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**

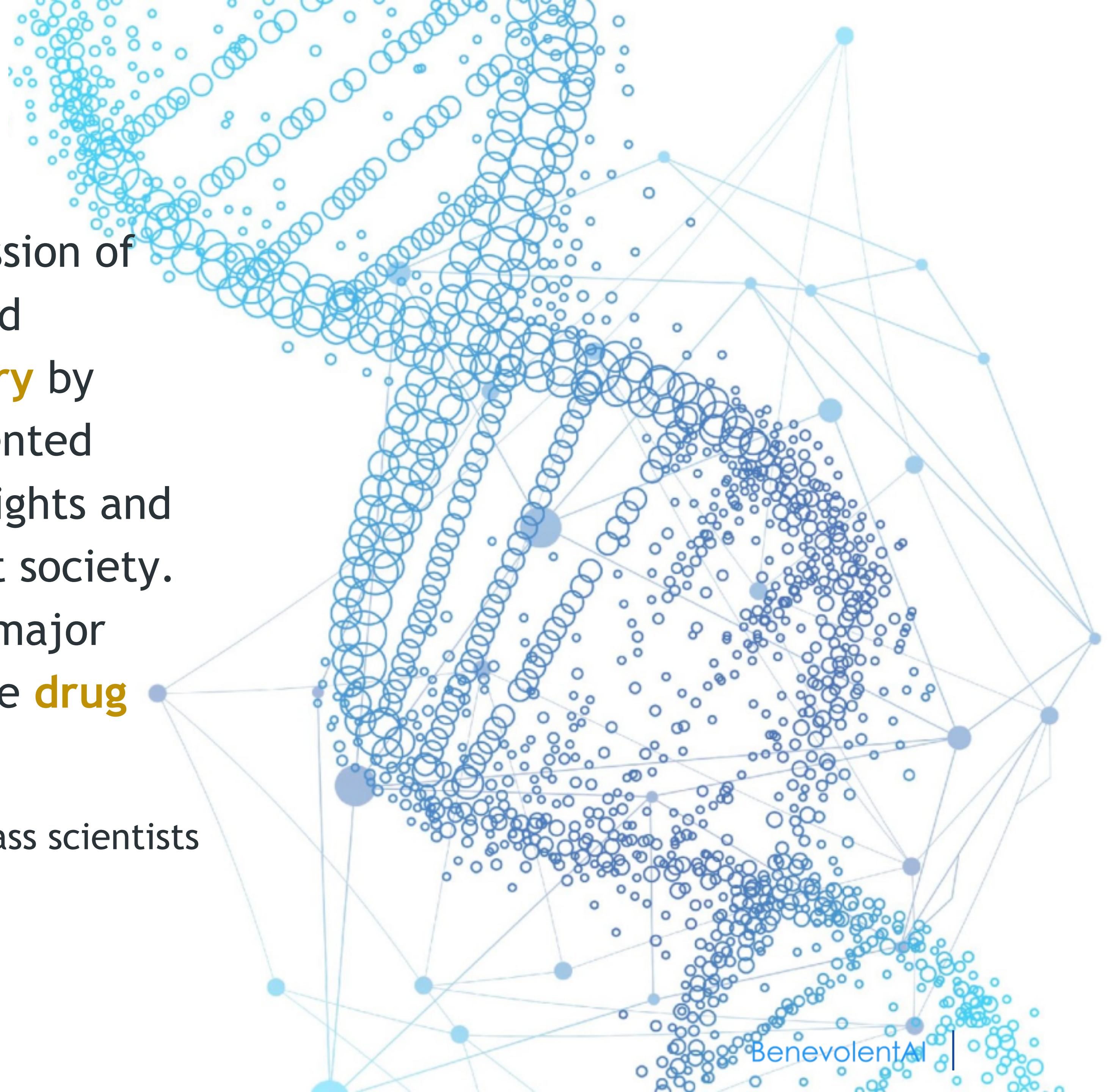
  - Reasonable benchmarks with strong baselines
  - Contentful contexts are often **event-centric**

- **+ Move towards Benevolent applications in real world!**

BenevolentAI

Founded in 2013, with the mission of harnessesing AI to enhance and **accelerate scientific discovery** by making sense of highly fragmented information to create new insights and usable knowledge that benefit society. Focused on **finding cures** for major diseases through disrupting the **drug discovery** industry.

$BN valuation, 85+ team of world-class scientists and technologists



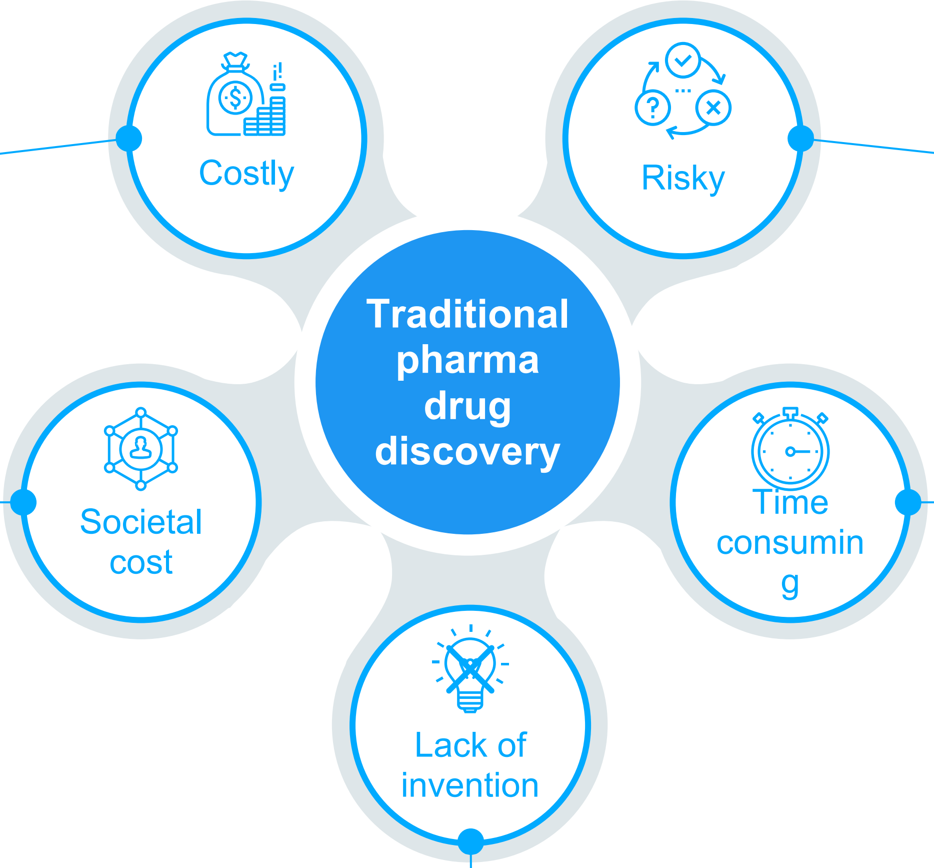
The existing drug discovery process is broken

$2.6Bn
cost to develop a new drug[1]

97%
of drug programmes fail[2]

Only 40%
known diseases are currently treatable[5]

12-15 years
from start to market[3]



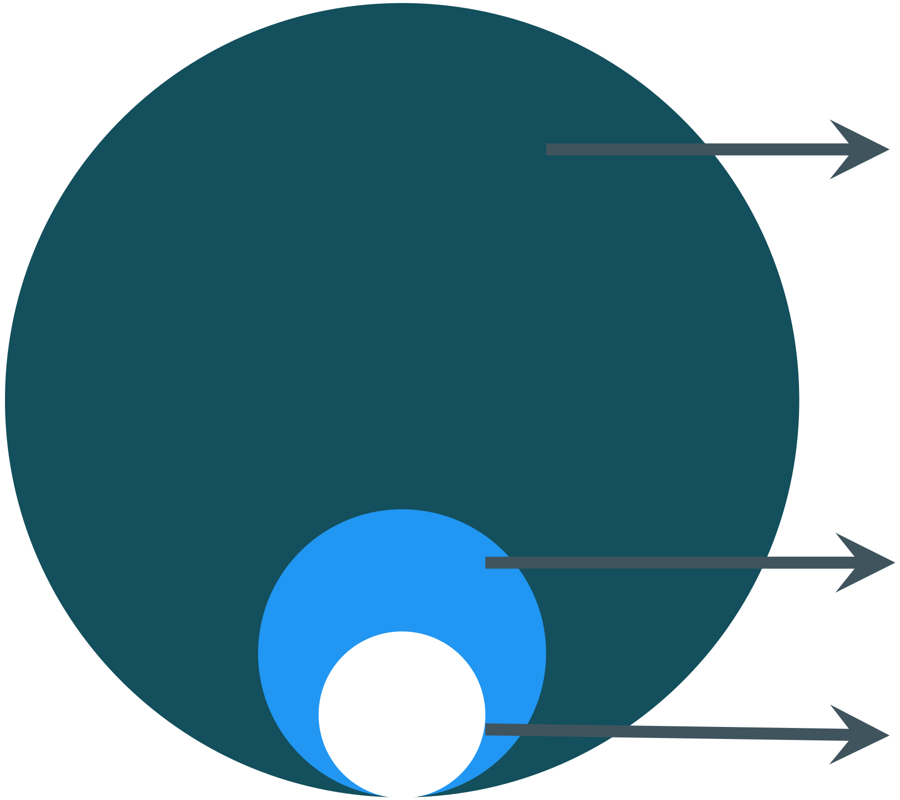
Extremely rare
finding cause of disease

[1] Tufts Center for the Study of Drug Development (PR Tufts CSDD 2014 Cost Study, November 2014).
[2] Biotechnology Innovation Organisation (Clinical development success rates, 2006-2015, February 2016).
[3] U S Food and Drug Administration (Drug Innovation - Novel Drug Approvals for 2016, January 2017).
[4] Department of Neuroscience, Uppsala University (Trends in the exploitation of novel drug targets, 2011).
[5] Data from BenevolentAI knowledge graph derived in part from MeSH, OMIM, Disease Ontology and Orphanet

Power of NLP and AI - Machine Reading and Reasoning

- **One scientific paper** is published **every 30 seconds,** which is impossible to be read manually!
- All biological databases combined are less than 5% of the available data

The BenevolentAI platform extracts facts and reasons from all relevant databases and literature, structured and unstructured



~90M full text patents

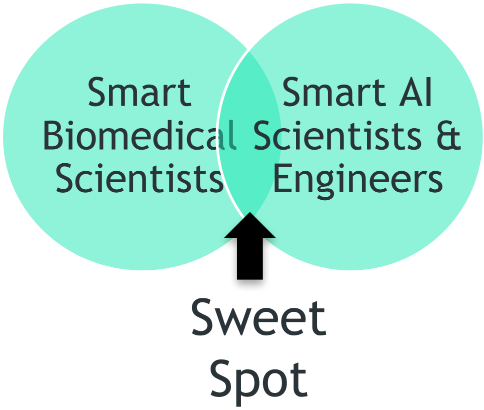
~30M scientific articles

~20% as structured data[1]

[1] https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

What’s Unique about BenevolentAI?

- AI and NLP researchers work directly with Drug Discovery scientists and Bioinformaticians every day
  - Whereas other AI groups often make partnerships with particular hospitals or research institutes, we work closely with biomedical scientists all the time
  - This allows us to develop technology that is laser-targeted to the need of the drug discovery scientists
  - Our research and technology improves quickly based on their constant feedback, building unique differentiation.
- We don’t need to hype AI!
  - We either find a cure for a major disease or not
    - This encourages being approach-agnostic
    - We are grounded in real-world performance as opposed to a narrow specific task



How can NLP help find cure for cancer?

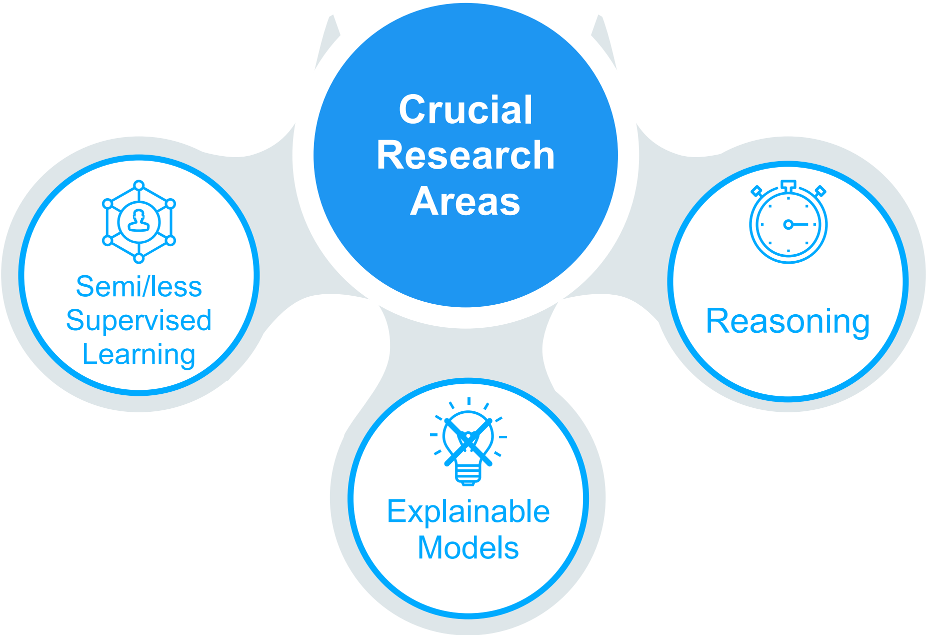
Regina Barzilay’s NAACL’16 Keynote:

- Reliable Information Extraction
- Interpretable Models

How can NLP help find cures for major diseases?

- **Reliable Semi-Supervised Information Extraction and Knowledge Acquisition**
- **Interpretable Reasoning**
  - We need to move towards “reasoning”, to discover beyond what is explicit
  - In our domain, the end-users (drug discovery scientists) always ask “why” for any prediction

How can NLP help find cures for major diseases?



From Public & Licensed data to Proprietary Knowledge

Unstructured Data

ANALYSIS

nature genetics

The support of human genetic evidence for approved drug indications

Matthew R Nelson[1], Hannah Tipney[2], Jeffery L Painter[1], Judong Shen[1], Paola Nicoletti[3], Yufeng Shen[3,4], Aris Floratos[3,4], Pak Chung Sham[5,6], Mulin Jun Li[6,7], Junwen Wang[6,7], Lon R Cardon[8], John C Whittaker[2] & Philippe Sanseau[2]

© 2015 Nature America, Inc. All rights reserved.

Over a quarter of drugs that enter clinical development fail because they are ineffective. Growing insight into genes that influence human disease may affect how drug targets and indications are selected. However, there is little guidance about how much weight should be given to genetic evidence in making these key decisions. To answer this question, we investigated how well the current archive of genetic evidence predicts drug mechanisms. We found that, among well-studied indications, the proportion of drug mechanisms with direct genetic support increases significantly across the drug development pipeline, from 2.0% at the preclinical stage to 8.2% among mechanisms for approved drugs, and varies dramatically among disease areas. We estimate that selecting genetically supported targets could double the success rate in clinical development. Therefore, using the growing wealth of human genetic data to select the best targets and indications should have a measurable impact on the successful development of new drugs.

Attrition is a major challenge in drug discovery and development, with more than half of clinical studies failing because of lack of efficacy[1–4]. The widespread failure of preclinical model systems to adequately predict efficacy in humans has led drug developers to look for other sources of evidence to inform decisions about which targets to pursue and for which indications (disease or reason for treatment for which a drug is approved). Since the completion of the Human Genome Project and the rise of genome-wide association studies (GWAS) and whole-genome and whole-exome sequencing studies, there has been rapid progress in identifying the genes that influence human health and disease[5]. These genetic insights can potentially transform the process of selecting the best drug targets and indications[6], the key decisions in drug discovery. There are several examples of genes associated with disease traits that have been proven to be effective drug targets. One canonical example is the target for statins, *HMGCR*, which has been associated with serum cholesterol levels[7]. Several other examples were recently highlighted for rheumatoid arthritis[8]. Such examples and the rapidly growing body of human genetic data led us to ask how much weight should be given to genetic associations when choosing which drug targets to pursue for a desired indication.

RESULTS

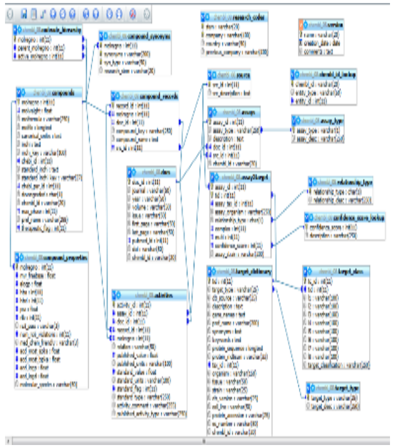
In this study, we go beyond previous work on drug repositioning[9] to investigate how well clinically successful drug mechanisms (the protein product modulated to elicit a clinical response) are predicted by known genetic associations and how that prediction may change across the drug development pipeline, from preclinical and clinical phases to launched drugs (Drug Approval Process; see URLs). An overview of the data sources, filtering and processing applied is provided in **Figure 1a**. To broadly capture statistically significant ($P \leq 1 \times 10^{-8}$) common variant genetic associations, we used GWASdb[10], which combines data from multiple sources, including the National Human Genome Research Institute (NHGRI) GWAS Catalog, the tables and supplementary materials of manuscripts archived in the NHGRI GWAS Catalog, and the database of Genotypes and Phenotypes (dbGaP), among others. To allow comparisons among all data sources, we manually mapped all traits to the most specific Medical Subject Heading (MeSH) terms applicable. Genetic variants were mapped to potential causal genes using a combination of linkage disequilibrium (LD), position, expression quantitative trait locus (eQTL) and epigenetic data (for example, see **Fig. 1b**). When we observed multiple possible variant-to-gene mappings, these were ranked on the overall strength of evidence. In the final data set, we had 18,566 genetic associations to 434 MeSH traits that mapped to 6,120 genes outside of the extended major histocompatibility complex (xMHC), with a total of 13,855 gene-trait combinations. Genes involved in rare, mendelian traits were derived from Online Mendelian Inheritance in Man (OMIM), providing a data set with 1,898 genes annotated to affect 2,145 traits with MeSH terms, for a total of 2,627 gene-trait combinations. The GWASdb and OMIM gene-MeSH pairs were largely non-overlapping, yielding a combined set of 16,459 gene-trait combinations (**Supplementary Fig. 1** and **Supplementary Table 1**).

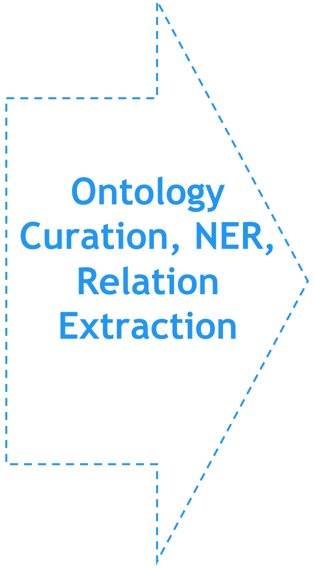
[1]Quantitative Sciences, GlaxoSmithKline, Research Triangle Park, North Carolina, USA. [2]Quantitative Sciences, GlaxoSmithKline, Stevenage, UK. [3]Department of Systems Biology, Columbia University, New York, New York, USA. [4]Department of Biomedical Informatics, Columbia University, New York, New York, USA. [5]State Key Laboratory of Brain and Cognitive Sciences, University of Hong Kong, Hong Kong, SAR, China. [6]Centre for Genomics Sciences, Li Ka Shing (LKS) Faculty of Medicine, University of Hong Kong, Hong Kong SAR, China. [7]Department of Biochemistry, LKS Faculty of Medicine, University of Hong Kong, Hong Kong SAR, China. [8]Alternative Discovery and Development, GlaxoSmithKline, Upper Merion, Pennsylvania, USA. Correspondence should be addressed to M.R.N. (matthew.r.nelson@gsk.com).

Received 23 September 2014; accepted 1 May 2015; published online 29 June 2015; doi:10.1038/ng.3314

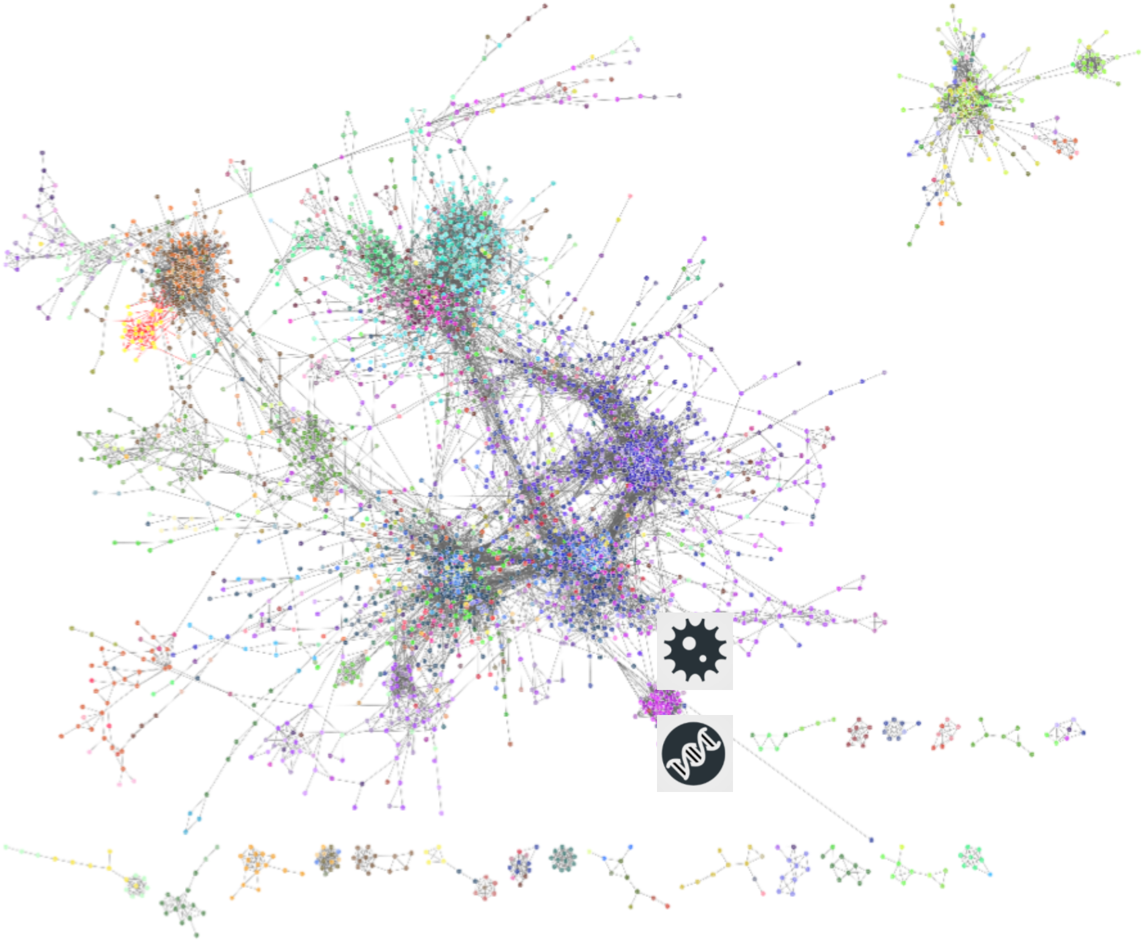
NATURE GENETICS ADVANCE ONLINE PUBLICATION 1

Structured Data


chembl_08.molecule_hierarchy
chembl_08.compound_synonyms
chembl_08.research_codes
chembl_08.version
chembl_08.compounds
chembl_08.compound_records
chembl_08.source
chembl_08.chembl_id_lookup
chembl_08.assays
chembl_08.assay_type
chembl_08.docs
chembl_08.assay2target
chembl_08.relationship_type
chembl_08.confidence_score_lookup
chembl_08.compound_properties
chembl_08.activities
chembl_08.target_dictionary
chembl_08.target_class
chembl_08.target_type

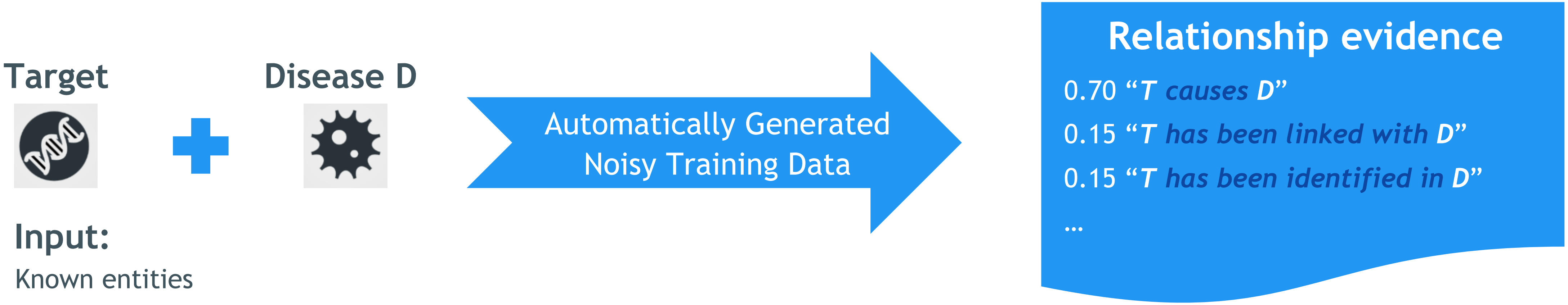


Rich proprietary knowledge graph



BenevolentAI

Distant Supervision for Relationship Extraction



Target

Disease D

Automatically Generated
Noisy Training Data

Relationship evidence

0.70 “*T causes D*”

0.15 “*T has been linked with D*”

0.15 “*T has been identified in D*”

…

Input:

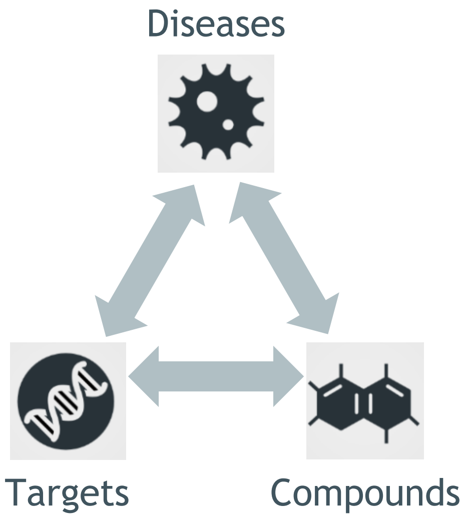
Known entities

Known relationships

Vast text corpora

Hypothesis Generation

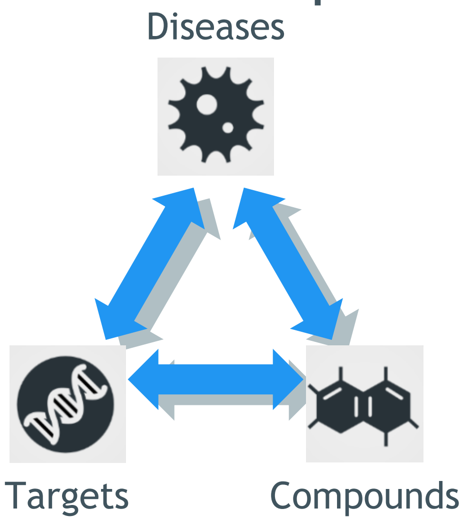
Known explicit relationships



Reasoning

- Tensor Factorization
- Path-inference models
- ...

Inferred implicit relationships



Input:

Known entities

Known relationships

Vast text corpora

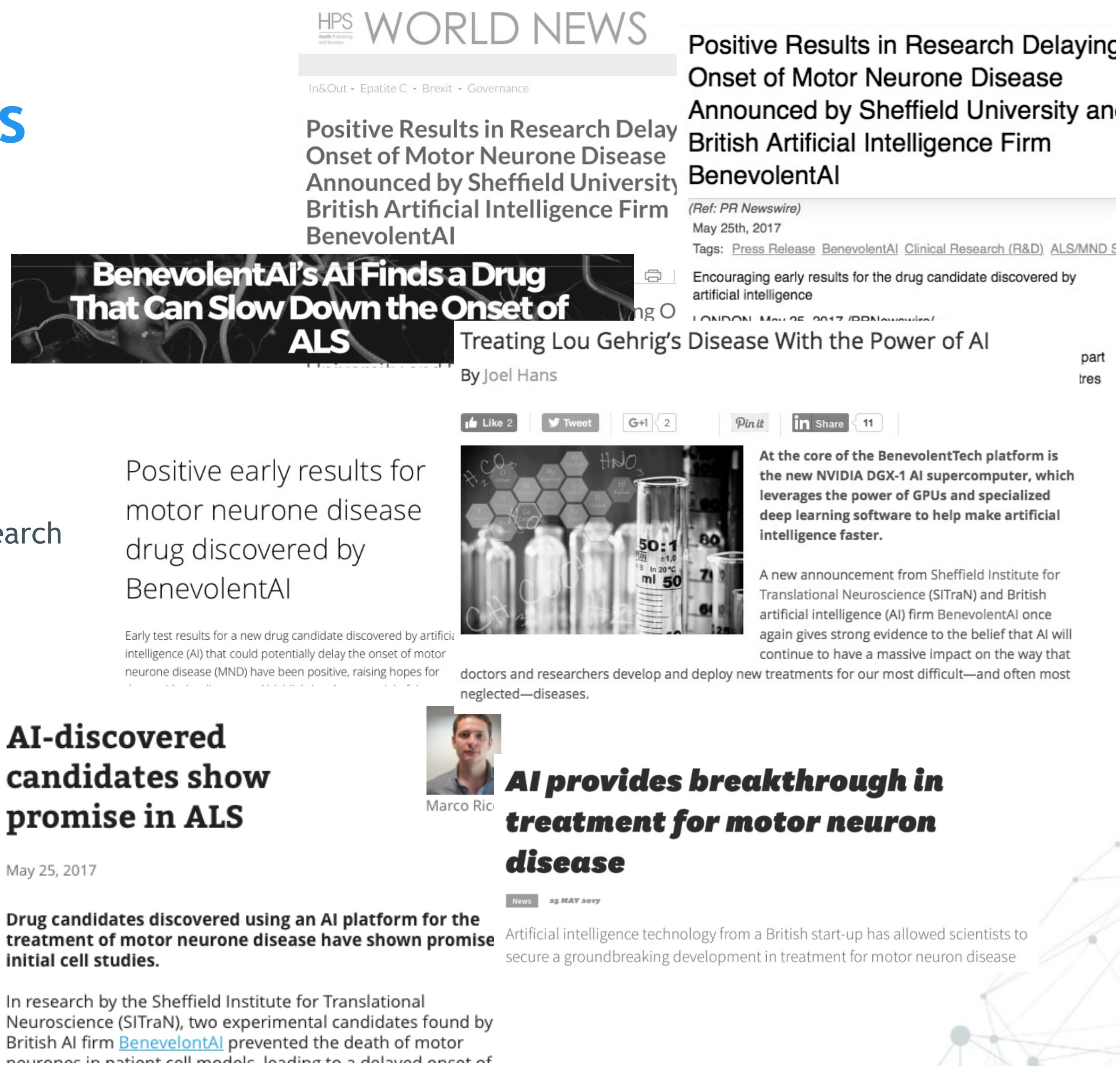
Output:

Inferred implicit relationships

BenevolentAI Discovers Potential cure for ALS

Our earliest study resulted in a stunning outcome

- 5 novel hypotheses generated in half a day
- Hypotheses validated by a world leading ALS research organisation (SITraN)
- Results showed delayed onset of ALS
- Progress announced by SITraN on 25 May 2017
- Clinical development 2018



HPS WORLD NEWS

In&Out - Epatite C - Brexit - Governance

Positive Results in Research Delay Onset of Motor Neurone Disease Announced by Sheffield University British Artificial Intelligence Firm BenevolentAI

Positive Results in Research Delaying Onset of Motor Neurone Disease Announced by Sheffield University an British Artificial Intelligence Firm BenevolentAI

(Ref: PR Newswire)

May 25th, 2017

Tags: Press Release BenevolentAI Clinical Research (R&D) ALS/MND

Encouraging early results for the drug candidate discovered by artificial intelligence

BenevolentAI's AI Finds a Drug That Can Slow Down the Onset of ALS

Treating Lou Gehrig's Disease With the Power of AI

By Joel Hans

At the core of the BenevolentTech platform is the new NVIDIA DGX-1 AI supercomputer, which leverages the power of GPUs and specialized deep learning software to help make artificial intelligence faster.

A new announcement from Sheffield Institute for Translational Neuroscience (SITraN) and British artificial intelligence (AI) firm BenevolentAI once again gives strong evidence to the belief that AI will continue to have a massive impact on the way that doctors and researchers develop and deploy new treatments for our most difficult—and often most neglected—diseases.

Positive early results for motor neurone disease drug discovered by BenevolentAI

Early test results for a new drug candidate discovered by artificia intelligence (AI) that could potentially delay the onset of motor neurone disease (MND) have been positive, raising hopes for

AI-discovered candidates show promise in ALS

May 25, 2017

Drug candidates discovered using an AI platform for the treatment of motor neurone disease have shown promise initial cell studies.

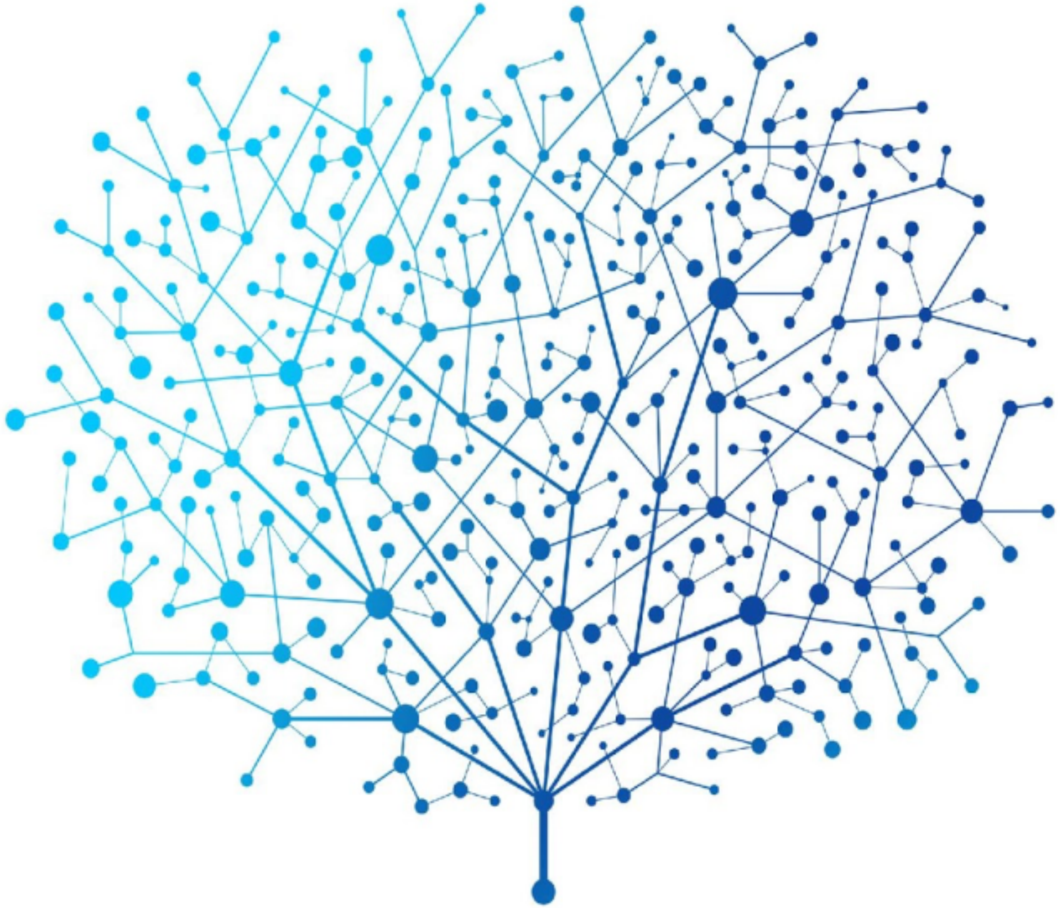
In research by the Sheffield Institute for Translational Neuroscience (SITraN), two experimental candidates found by British AI firm BenevelontAI prevented the death of motor

Marco Ric

AI provides breakthrough in treatment for motor neuron disease

News 25 MAY 2017

Artificial intelligence technology from a British start-up has allowed scientists to secure a groundbreaking development in treatment for motor neuron disease



We are hiring in NYC!

* Full-time NLP/ML Researchers

* Summer 2018 Research Internships

Get in touch!

nasrin.m@benevolent.ai

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

Paidi Creed, Aaron Sim, Jiajie Zhang

& Hoifung Poon and Regina Barzilay for inspiring NLP researchers to work on biomedicine!

1850 MELIORA UNIVERSITY of ROCHESTER

Microsoft Research

NAVY

Facebook AI Research

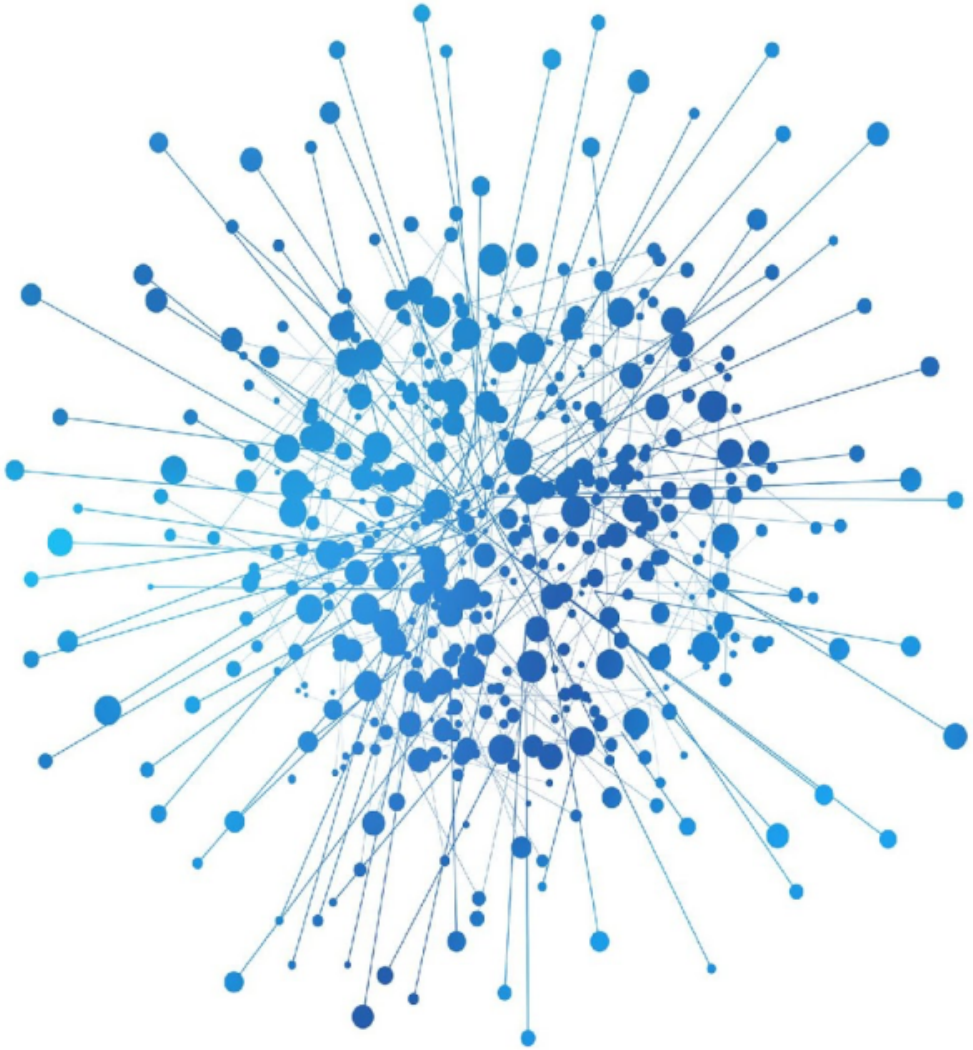
Carnegie Mellon University

Georgia Tech

Google

DeepMind

BenevolentAI



Thanks for Listening

Any Questions?