NAACL 2016:
June 2016

A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories

Nasrin Mostafazadeh
University of Rochester

In Collaboration with:

Nate Chambers (USNA), Pushmeet Kohli (MSR), Devi Parikh (VTech), Dhruv Batra (VTech), Lucy Vanderwende (MSR), Xiaodong He (MSR), James Allen (Rochester)

Story Understanding and Story Generation

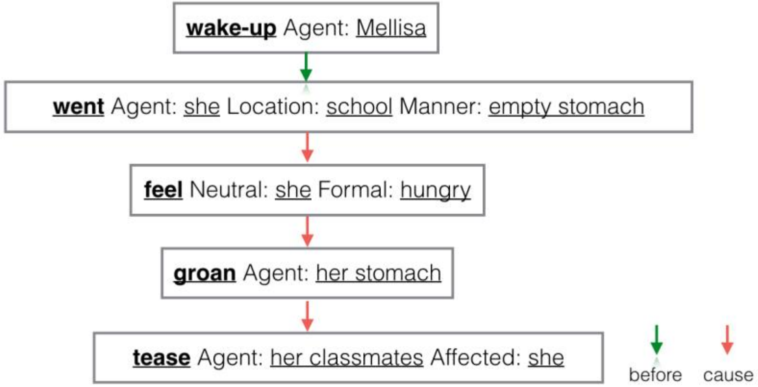
- An extremely challenging task in AI (Charniak 1972; Turner, 1994; Schubert and Hwang, 2000)

- **Perhaps the biggest challenge**: having commonsense knowledge for the interpretation of narrative events.

- How to provide commonsense knowledge regarding daily events to machines?

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

Story: Mellisa woke up quite late. She went to school on empty stomach. She felt hungry soon. Her stomach groaned. She was teased by her classmates.



wake-up Agent: Mellisa

went Agent: she Location: school Manner: empty stomach

feel Neutral: she Formal: hungry

groan Agent: her stomach

tease Agent: her classmates Affected: she

Overview of This Talk

- Introduce a new corpus for story modeling, called **ROCStories**.
- Introduce a new evaluation framework for benchmarking progress on story modeling and narrative structure learning, called **Story Cloze Test**.

What Is a Story?

- We define a narrative or story as follows:

“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?

- People had tried newswire. However, there is no much commonsense knowledge about daily events in news articles!
  - **One observation**: Personal stories from Weblogs are great sources of commonsense causal information (Gordon and Swanson, 2009).

ICWSM 2011 Spinn3r Dataset (Burton et al., 2009): tens of millions of non-spam weblog entries, aggregated by Spinn3r.com for research purpose.

“I cracked the egg into the bowl and then I saw it,
yeah a baby chicken was in right inside the egg that
was going to be our breakfast. I felt like I might be
sick, but the
rest of my family found this to be very
interesting!…FIGURE…You see, that’s what I’m
talking about.”

Teasing out useful information from noisy blog entries is a problem of its own!

ROCStories!

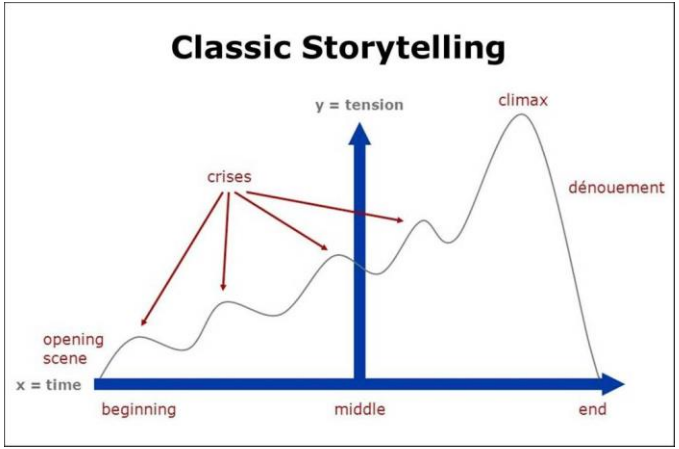
ROCStories: Short Commonsense Stories

- A collection of high quality **short five-sentence stories** with **their titles** authored by **hundreds of crowd workers**.
  - The five-sentence length gives enough context to the story, without giving room for sidetracking to less important or irrelevant information throughout the story.
  - Characteristics:

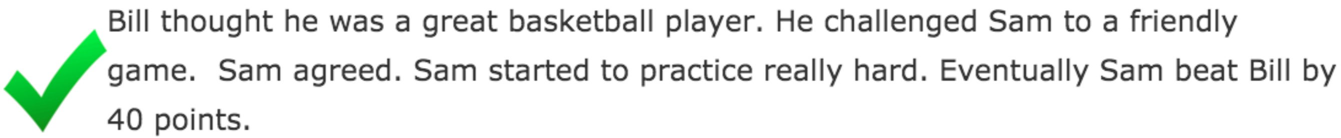
    (1) Is realistic

    (2) Has a **specific beginning and ending**, where **something happens in between**

    (3) Has nothing irrelevant or redundant to the core story



An Example Story



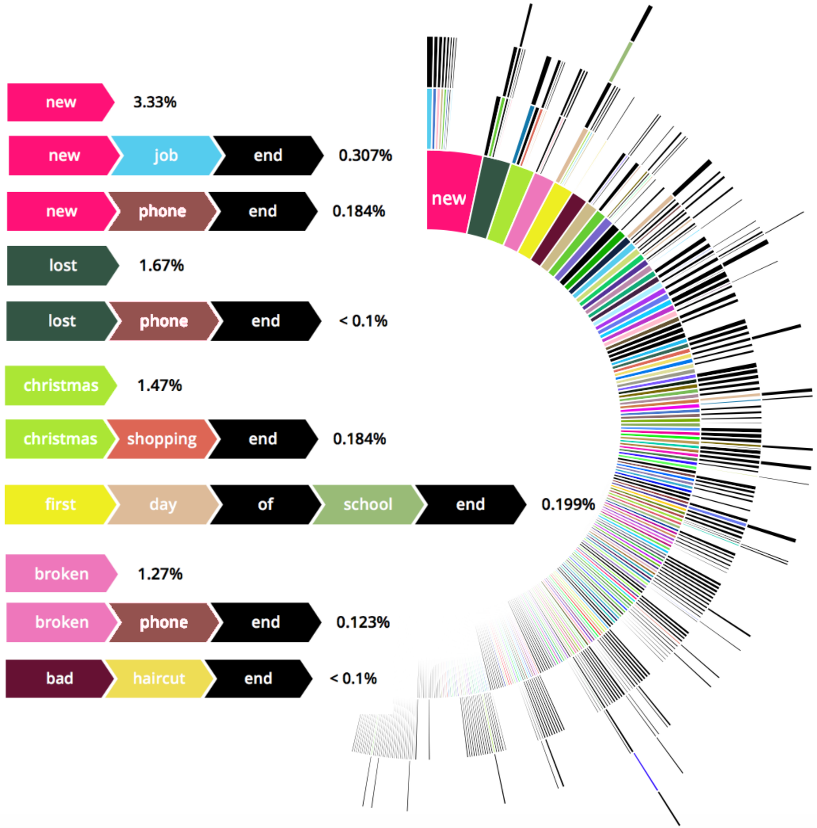
✓ Bill thought he was a great basketball player. He challenged Sam to a friendly game. Sam agreed. Sam started to practice really hard. Eventually Sam beat Bill by 40 points.

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

Statistics

- We’ve collected **49,255** stories so far.
- Total number of Turkers participated: **932**
- Average number of HITs done by one Turker: **52.84**
- Max number of HITs done by one Turker: **3057**

Data Quality: Title N-gram Distribution



new 3.33%

new job end 0.307%

new phone end 0.184%

lost 1.67%

lost phone end < 0.1%

christmas 1.47%

christmas shopping end 0.184%

first day of school end 0.199%

broken 1.27%

broken phone end 0.123%

bad haircut end < 0.1%

Data Quality: Temporal Analysis

- How well can human rearrange the shuffled sequence of sentences in our five-sentence stories?
    - Of 250 shuffled stories, % perfectly ordered: **95.2**
    - Of 250 placements for each position, % correct in each position:

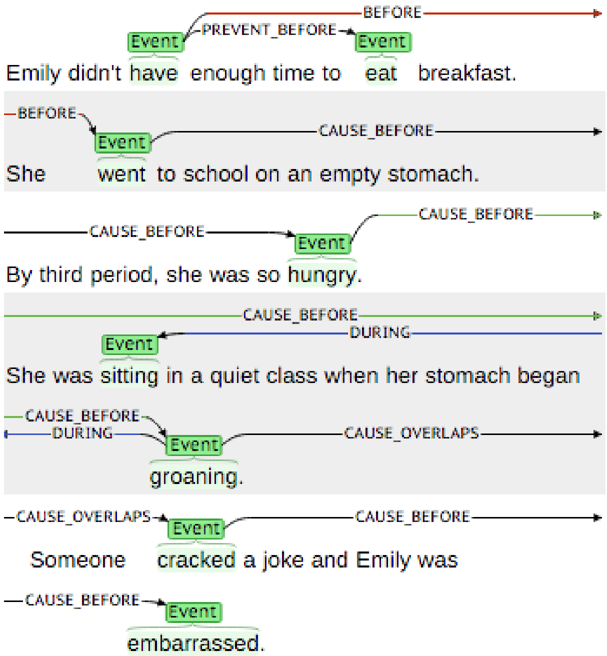
97.2 91.8 91.3 93.6 **97.8**

This further verifies the
richness of our corpus in terms of logical relation
between events!

Semantic Analysis of the Event Structures of ROCStories

(To be Presented at NAACL Event Workshop 2016, this Friday!)

- We developed CaTeRS: Causal and Temporal Relation Scheme for Semantic Annotation of Event Structures



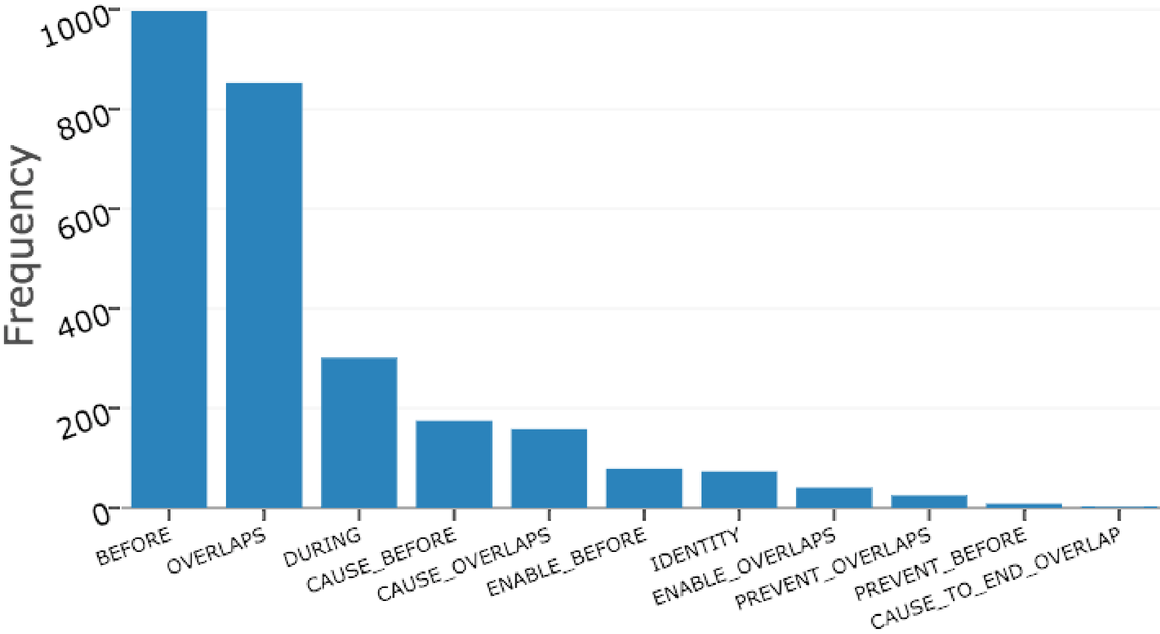
Emily didn't have enough time to eat breakfast.

She went to school on an empty stomach.

By third period, she was so hungry.

She was sitting in a quiet class when her stomach began groaning.

Someone cracked a joke and Emily was embarrassed.



Inter-annotator agreement Kappa: event entities 0.91, semantic links: 0.51

Evaluation

How to do automatic evaluation on story understanding/narrative structure learning?

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

Our Idea: Story Cloze Test!

- **Goal**: Design a new evaluation schema for story understanding and narrative structure learning.
- **State-of-the-art Evaluation:** Narrative cloze test (Chambers and Jurafsky, 2008) where a system predicts a held-out event given a sequence of observed ones.

  {X threw, pulled X, told X, ???, X completed}

  - Not meant to be solvable by human
  - Not foolproof (Pichotta and Mooney, 2014 and Rudinger et al, 2015)
  - No fixed human-verified test set shared in the community
- **Proposed Task:** Given a context of four sentences, predict the ending of the story.
  - Select from the 'right' and 'wrong' ending choices which are crowdsourced.

An Examples 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 Endings by Two Turkers**:
  - He proposed to Sheryl and she said Yes!
  - Tom asked Sheryl to marry him.
- **Wrong Endings by Two Turkers**:
  - He wiped mud off of his boot.
  - Tom tied his shoe and left Sheryl.

pinterest.com

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

Is this Evaluation Foolproof?!

Each model should choose the right ending given the context …

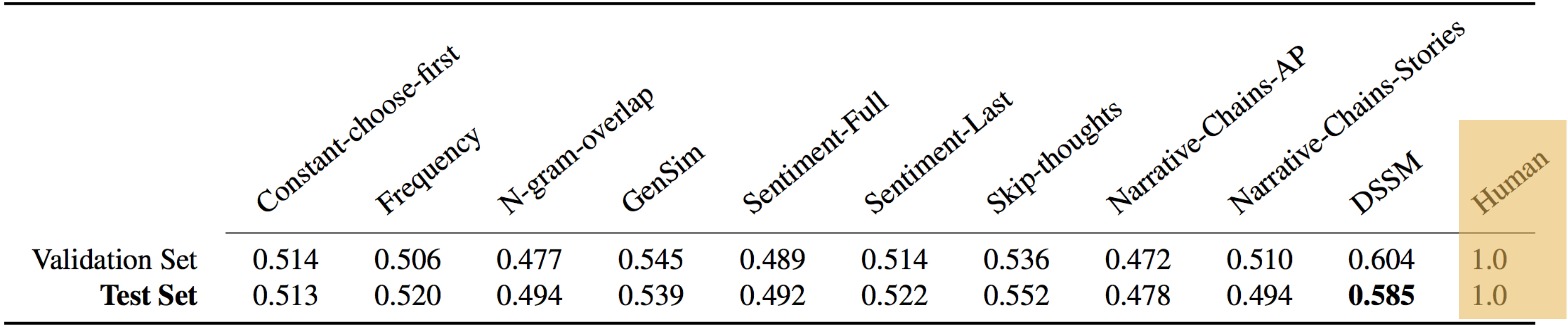
*** Cheap Tricks**

1. **Frequency** (discard the context): Choose the ending with higher (search engine hits) frequency of the main event.
2. **N-gram overlap**: Choose the ending with higher n-gram overlap with the context, computed using Smoothed-BLEU metric.
3. **Average Word2Vec (GenSim)**: Choose the ending with closer average word2vec to the average word2vec of the four-sentences context (this is basically an enhanced 'word overlap' baseline, which accounts for synonyms).
4. **Sentiment Match**: Choose the ending that matches the sentiment of the four-sentences context (Full) or the fourth-sentence (Last).

* **Models**

5. **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.
6. **Narrative Chains**: (Chambers and Jurafsky, 2009) trained once on AP portion of Gigawords Corpus (-AP) and then on ROCStories (-Stories), this model computes PMI between event pairs and chooses the ending with the highest total PMI in the chain of context events.
7. **Deep Structured Semantic Model (DSSM)**: MSR Sentence2Vec model (Huang et al., 2013), according to which you can choose the option that has a closer embedding to the four-sentences context.

Results on Story Cloze Test



	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	0.604	1.0
Test Set	0.513	0.520	0.494	0.539	0.492	0.522	0.552	0.478	0.494	**0.585**	1.0

Story Cloze Test is a great new framework for evaluating language understanding, specifically narrative structure learning and story understanding!

Why Story Cloze Test is an Outstanding Evaluation for Broad-coverage NLU ?!

- Foolproof!
- Human performs 100%.
- There is a wide enough gap from the state-of-the-art to human performance (42% gap), so plenty of room for research!

Conclusion

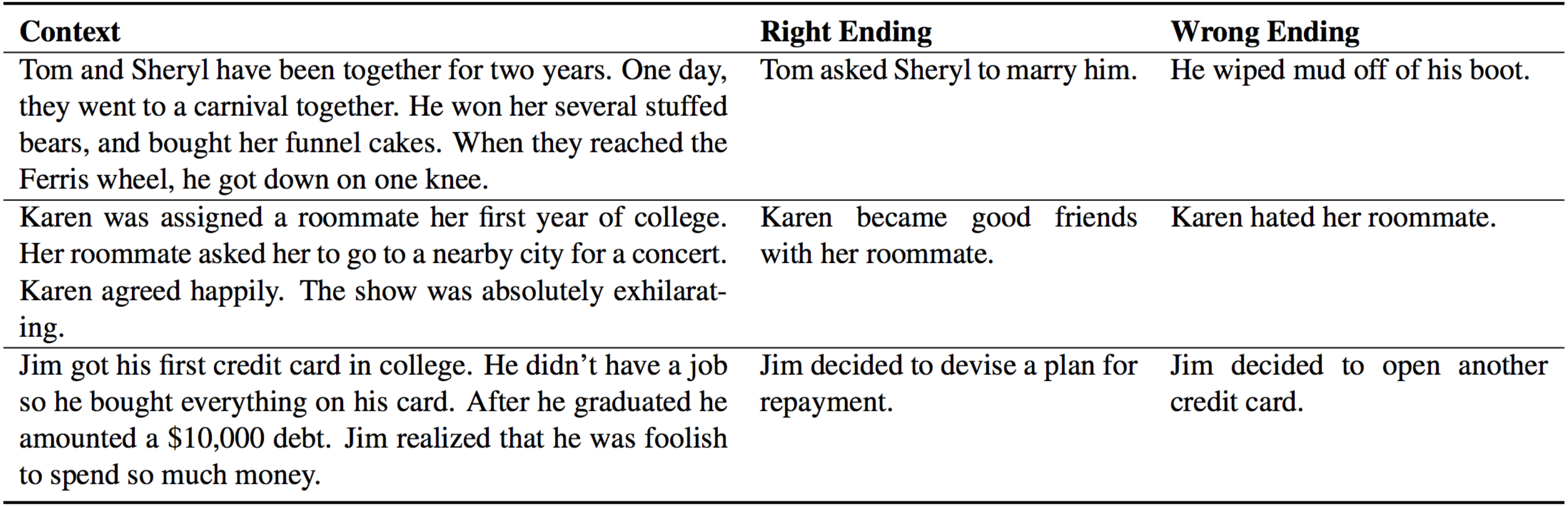
- Releasing a new dataset of ~50,000 short commonsense stories, called ROCStories, which can be used for any story generation or narrative structure learning purposes.
- Introducing a new framework for evaluating Story Generation models as well as Narrative Structure Learners, called Story Cloze Test.
- We will set up Story Cloze Test as a challenge on CodaLab, please stay tuned to follow the leaderboard.
- Dataset can be found here: http://cs.rochester.edu/nlp/rocstories/

Thanks a lot for your Attention

Any Questions?

Backup Slides…

Some Example Story Cloze Tests



Context	Right Ending	Wrong Ending
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 on one knee.	Tom asked Sheryl to marry him.	He wiped mud off of his boot.
Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating.	Karen became good friends with her roommate.	Karen hated her roommate.
Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money.	Jim decided to devise a plan for repayment.	Jim decided to open another credit card.

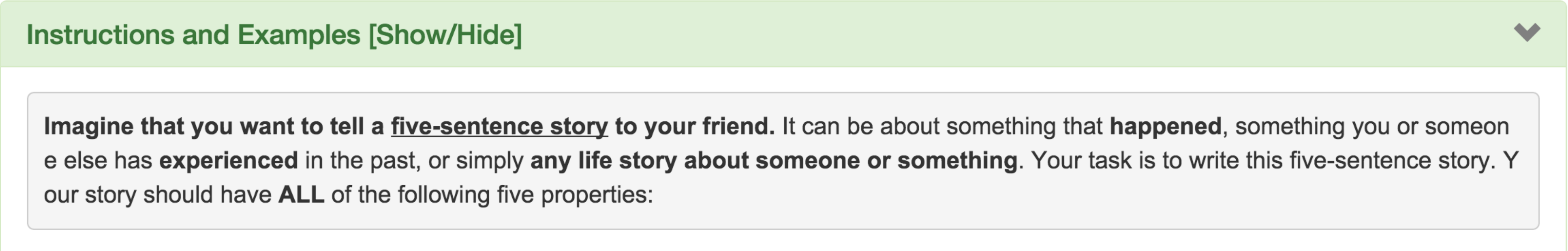
Some Example Stories

- Jennifer has a big exam tomorrow. She got so stressed, she pulled an all-nighter. She went into class the next day, weary as can be. Her teacher stated that the test is postponed for next week. Jennifer felt bittersweet about it

- James woke up in the middle of the night to find his bed shaking. All the things in his room were shaking furiously. He was scared, but then suddenly it stopped. He got up and checked the news. He found out there had been an earthquake!

- Amy's friend Beth was having a baby. Amy called her friends to come to a surprise baby shower. Everyone waited quietly for Beth to enter. She was shocked when she saw everyone. Beth was so happy since she thought no one cared to give her a shower.

Crowd-sourcing on AMT: They key is the prompt!

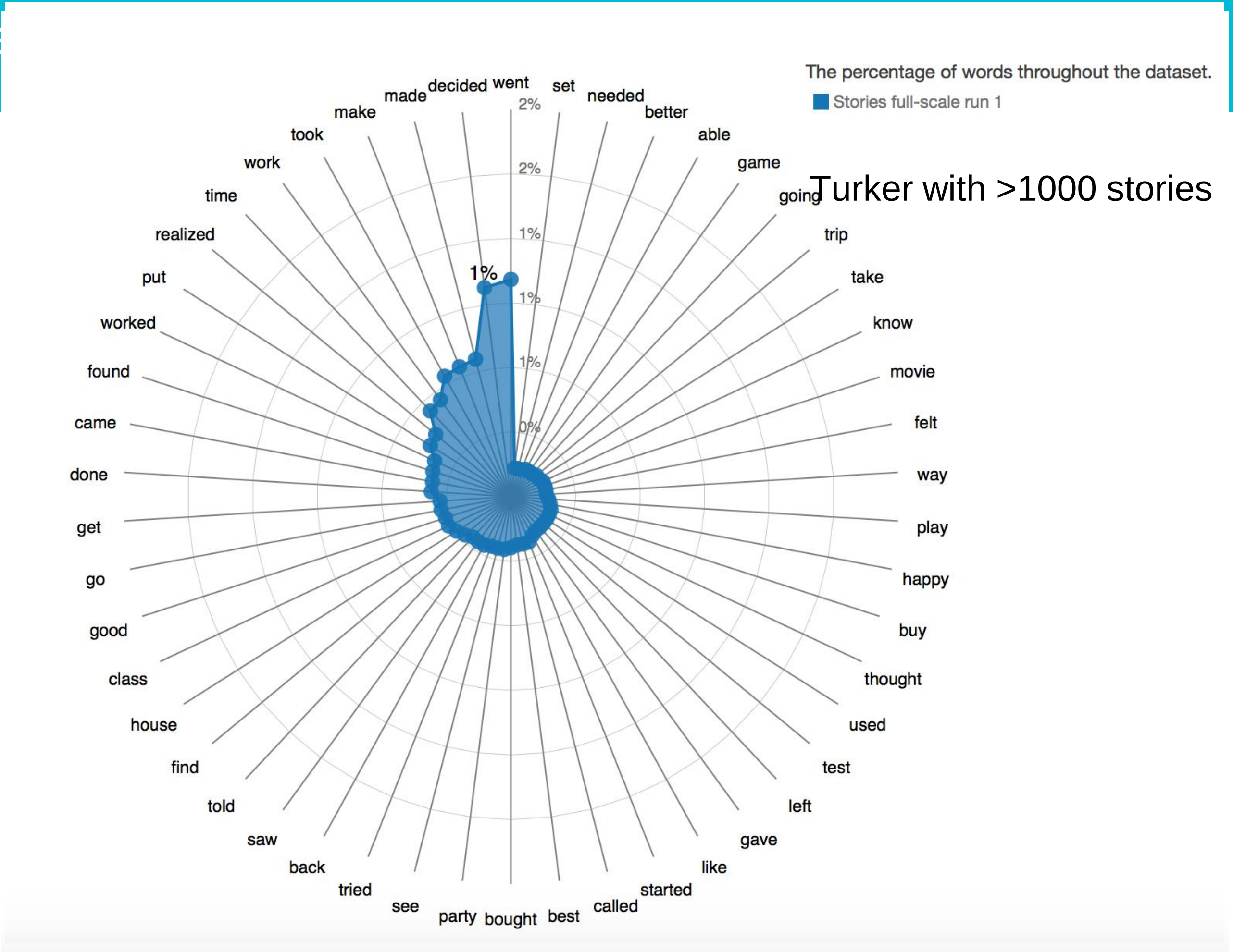


Instructions and Examples [Show/Hide]

Imagine that you want to tell a five-sentence story to your friend. It can be about something that **happened**, something you or someone else has **experienced** in the past, or simply **any life story about someone or something**. Your task is to write this five-sentence story. Your story should have **ALL** of the following five properties:

1. Your story should be entirely **realistic.**
2. Your story should read like a coherent story, with a specific beginning and end, where something happens in between.
3. Each sentence in your story should be **logically related to the next sentence** and be about the characters of the story.

Dataset Bia



The percentage of words throughout the dataset.

Turker with >1000 stories

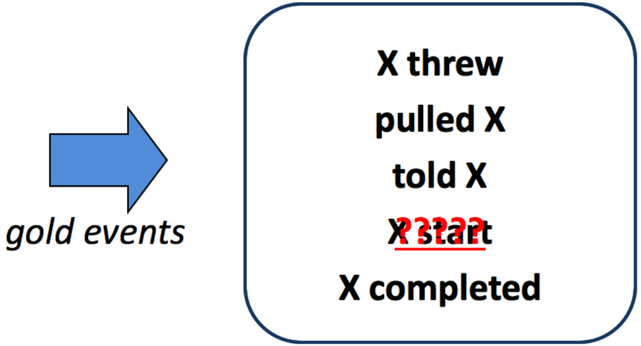
Narrative Cloze Test

- The state-of-the-art Evaluation Framework:
  - Narrative Cloze Test: Predict the missing event, given a set of observed events.

What was the original goal of this evaluation?

1. "comparative measure to evaluate narrative knowledge"
2. **"not meant to be solvable by humans"**

McCann threw two interceptions early… Toledo pulled **McCann** aside and told **him he**'d start… **McCann** quickly completed his first two passes…



Narrative Cloze Test & Issues!

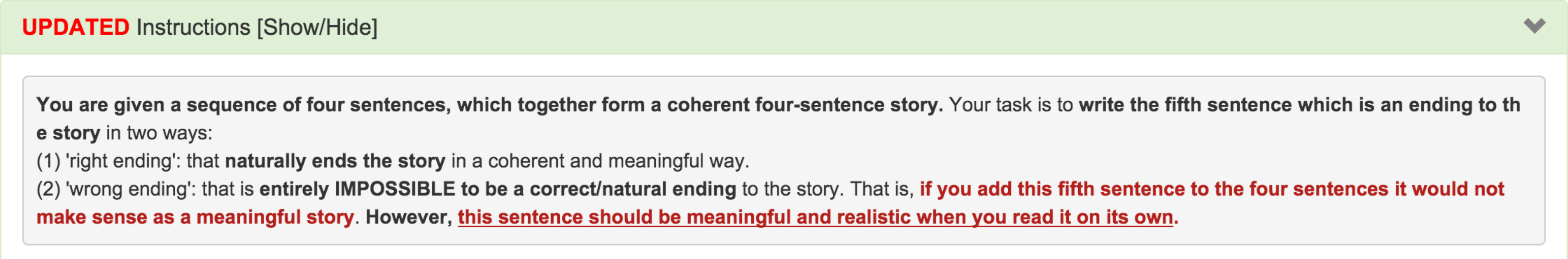
Do you need narrative schemas to perform well?

- Current Narrative Cloze Tests are auto-generated from parses and coreference systems.
    - The event chains aren't manually verified as gold (as the original Narrative Cloze did) Jans et al., (2012) Pichotta and Mooney (2014) Rudinger et al. (2015)
- **As with all things in NLP, the community optimized evaluation performance, and not the big picture goal!**
    - Pichotta and Mooney (2014) showed that simply predicting the most frequent unigram is an extremely high baseline!
    - Language modeling is better than PMI on the Narrative. Rudinger et al (2015)

This is a major problem in community:

- Do we even care about predicting "X said"?

Crowdsourcing Story Cloze Test! —> Scalable!



UPDATED Instructions [Show/Hide]

You are given a sequence of four sentences, which together form a coherent four-sentence story. Your task is to **write the fifth sentence which is an ending to the story** in two ways:

(1) 'right ending': that **naturally ends the story** in a coherent and meaningful way.

(2) 'wrong ending': that is **entirely IMPOSSIBLE to be a correct/natural ending** to the story. That is, **if you add this fifth sentence to the four sentences it would not make sense as a meaningful story**. **However, this sentence should be meaningful and realistic when you read it on its own.**

1. The sentence should **follow up the story by sharing at least one of the characters of the story**.
2. The sentence should be entirely **realistic and meaningful when read in isolation by itself.**
   - ~~'Mary gave birth to a Raccon' or 'Mary and John walked on Mars'.~~

Applications

- Contain thousands of eventful stories, so it can be used for any story generation purposes.

- Capture a lot of commonsensical causal and temporal relations between world events, which can be used for various knowledge extraction and narrative structure learning purposes.