

Advertisement for ACL Workshops

Workshop on Narrative Understanding, Storylines, and Events (NUSE)

We solicit papers related to narrative understanding and all aspects of event and storyline analysis, story generation, and relationships between events and storylines that present new datasets, systems and methods, and evaluation methodologies.

Workshop on Neural Generation and Translation

Topics of interest include neural models for generation, dialogue, summarization, and simplification; analysis of the problems and opportunities of neural models for all of these tasks; handling resource-limited domains; and more.

Submission Deadline: April 6
Papers are 4 or 8 pages.

Natural Language Reasoning

Daphne Ippolito
Chris Callison-Burch

Examples of reasoning

Counting

Amy has five apples. She gives two to John. How many apples for Amy have?

Examples of reasoning

Counting

Amy has five apples. She gives two to John. How many apples for Amy have?

Translation

When translating the “telephone is working” and “the electrician is working” into German, the translations of “working” should be different.

Examples of reasoning

Counting

Amy has five apples. She gives two to John. How many apples for Amy have?

Translation

When translating the “telephone is working” and “the electrician is working” into German, the translations of “working” should be different.

Taxonomic Reasoning

If Fido is a dog and dogs are mammals, then Fido is a mammal.
If mammals are furry, then Fido is furry.

Examples of reasoning

Temporal Reasoning

If one knows that Mozart was born earlier and died younger than Beethoven, one can infer that Mozart died earlier than Beethoven.

Examples of reasoning

Temporal Reasoning

If one knows that Mozart was born earlier and died younger than Beethoven, one can infer that Mozart died earlier than Beethoven.

Common knowledge

These are often facts so basic, they aren't even written down.

“It takes a 10 minutes, not 10 days to make a cup of coffee. “

“Goats have two horn while unicorns only have one.”

Examples of reasoning

Temporal Reasoning

If one knows that Mozart was born earlier and died younger than Beethoven, one can infer that Mozart died earlier than Beethoven.

Common knowledge

These are often facts so basic, they aren't even written down.

“It takes a 10 minutes, not 10 days to make a cup of coffee. “

“Goats have two horn while unicorns only have one.” “Milk is white.”

World Knowledge

These are the kind of facts that appear in Wikipedia or other knowledge bases.

“The capital of Pennsylvania is Harrisburg.”

“Barack Obama was the 44th president of the United States.”

Tasks to evaluate language reasoning

Semantic Role Labeling

Relation Extraction

Event Factuality

Named Entity Recognition

Word Sense Disambiguation

Reference Resolution

Grammaticality

Lexicosyntactic Inference

Sentiment Analysis

Figurative Language

Sentence Similarity

Paraphrase

Sentence Completion

Textual Entailment

Question Answering

List from “Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches”

Tasks to evaluate language reasoning

Semantic Role Labeling

Relation Extraction

Event Factuality

Named Entity Recognition

Word Sense Disambiguation

Reference Resolution

Grammaticality

Lexicosyntactic Inference

Sentiment Analysis

Figurative Language

Sentence Similarity

Paraphrase

Sentence Completion

Textual Entailment

Question Answering

List from “Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches”

Tasks to evaluate language reasoning

Semantic Role Labeling

Relation Extraction

Event Factuality

Named Entity Recognition

Word Sense Disambiguation

Reference Resolution

Grammaticality

Lexicosyntactic Inference

Sentiment Analysis

Figurative Language

Sentence Similarity

Paraphrase

Sentence Completion

Story Completion

Textual Entailment

Question Answering

List from “Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches”

ROCStories Evaluation Task

Context	Ending 1	Ending 2
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 open another credit card.	Jim decided to devise a plan for repayment.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She didn't want her phone anymore.	She found her phone in the car.

ROCStories Evaluation Task

Context	Ending 1	Ending 2
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 open another credit card.	Jim decided to devise a plan for repayment.
Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside.	She didn't want her phone anymore.	She found her phone in the car.

ROCStories Dataset

Title	Five-sentence Story
The Test	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.
The Hurricane	Morgan and her family lived in Florida. They heard a hurricane was coming. They decided to evacuate to a relative's house. They arrived and learned from the news that it was a terrible storm. They felt lucky they had evacuated when they did.
Spaghetti Sauce	Tina made spaghetti for her boyfriend. It took a lot of work, but she was very proud. Her boyfriend ate the whole plate and said it was good. Tina tried it herself, and realized it was disgusting. She was touched that he pretended it was good to spare her feelings.

ROCStories - about the dataset

Amazon Mechanical Turk workers were asked to write 5-sentence long “everyday life stories” with a clear beginning and end with something happening in between.

The stories are intended to be short and simple; and common sense is necessary to make a good prediction of which 5th sentence is more likely.

Limitations

Sentences don't resemble most other natural language datasets (vocabulary is much simpler and sentences are shorter than other corpora).

Stories reflect the humans that wrote them.

The man made a lewd joke. A woman called him childish. The man wanted to look more adult. He started speaking in a lower voice. It made the woman respect him.

Harriet's bff's birthday is today. She wanted to get her bff something nice. Harriet decided to get flowers for her best friend. Harriet poked her own eye out while trimming her bff's flowers. Her bff was excited about the flowers as she drove to the hospital.

One of my daughter's high school friends got addicted to oxycontin. She was 19 and had dropped out of college. She was so addicted she stole money from her mom and aunt. She checked into a rehab center under threat of arrest. She has been clean for five years now.

Flora had a child that she adored. Flora was an alcoholic so she lost custody. She really wanted to see her child. She decided to go pick her child up. Flora kidnapped the child.

	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

	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

Rank	CodaLab Id	Model	ROCStories	Pre-trained Embeddings	Other Resources	Accuracy
1	msap	Logistic regression	Spring 2016, Winter 2017	—	NLTK Tokenizer, Spacy POS tagger	0.752
2	cogcomp	Logistic regression	Spring 2016, Winter 2017	Word2Vec	UIUC NLP pipeline, FrameNet, two sentiment lexicons	0.744
3	tbmihaylov	LSTM	—	Word2Vec	—	0.728
4	ukp	BiLSTM	Spring 2016, Winter 2017	GloVe	Stanford CoreNLP, DKPro TC	0.717
5	acoli	SVM	—	GloVe, Word2Vec	—	0.700
6	roemmele	RNN	Spring 2016, Winter 2017	Skip-Thought	—	0.672
7	mflor	Rule-based	—	—	VADER sentiment lexicon, Gigaword corpus PMI scores	0.621
8	Pranav.Goel	Logistic regression	Spring 2016, Winter 2017	Word2Vec	VADER sentiment lexicon, SICK data set	0.604
9	ROC_NLP (baseline)	DSSM	Spring 2016, Winter 2017	—	—	0.595

	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

Rank	CodaLab Id	Model	ROCStories	Pre-trained Embeddings	Other Resources	Accuracy
1	msap	Logistic regression	Spring 2016, Winter 2017	—	NLTK Tokenizer, Spacy POS tagger	0.752
2	cogcomp	Logistic regression	Spring 2016, Winter 2017	Word2Vec	UIUC NLP pipeline, FrameNet, two sentiment lexicons	0.744
3	tbmihaylov	LSTM	—	Word2Vec	—	0.728
4	ukp	BiLSTM	Spring 2016, Winter 2017	GloVe	Stanford CoreNLP, DKPro TC	0.717
5	acoli	SVM	—	GloVe, Word2Vec	—	0.700
6	roemmele	RNN	Spring 2016, Winter 2017	Skip-Thought	—	0.672
7	mflor	Rule-based	—	—	VADER sentiment lexicon, Gigaword corpus PMI scores	0.621
8	Pranav.Goel	Logistic regression	Spring 2016, Winter 2017	Word2Vec	VADER sentiment lexicon, SICK data set	0.604
9	ROC_NLP (baseline)	DSSM	Spring 2016, Winter 2017	—	—	0.595

Methods	Accuracy (%)
BERT _{BASE} (multilingual, uncased)	75.9
BERT _{BASE} (multilingual, cased)	80.2
BERT _{BASE} (monolingual, cased)	87.4
BERT _{BASE} (monolingual, uncased)	88.1
BERT _{LARGE} (monolingual, uncased)	89.2
BERT _{LARGE} (monolingual, cased)	90.0

SWAG

On stage, a woman takes a seat at the piano. She

- a) sits on a bench as her sister plays with the doll.
- b) smiles with someone as the music plays.
- c) is in the crowd, watching the dancers.
- d) nervously sets her fingers on the keys.

SWAG

On stage, a woman takes a seat at the piano. She

- a) sits on a bench as her sister plays with the doll.
- b) smiles with someone as the music plays.
- c) is in the crowd, watching the dancers.
- d) nervously sets her fingers on the keys.**

SWAG

A girl is going across a set of monkey bars. She

- a) jumps up across the monkey bars.
- b) struggles onto the monkey bars to grab her head.
- c) gets to the end and stands on a wooden plank.
- d) jumps up and does a back flip.

SWAG

A girl is going across a set of monkey bars. She

- a) jumps up across the monkey bars.
- b) struggles onto the monkey bars to grab her head.
- c) gets to the end and stands on a wooden plank.**
- d) jumps up and does a back flip.

SWAG

The woman is now blow drying the dog. The dog

- a) is placed in the kennel next to a woman's feet.
- b) washes her face with the shampoo.
- c) walks into frame and walks towards the dog.
- d) tried to cut her face, so she is trying to do something very close to her face.

SWAG

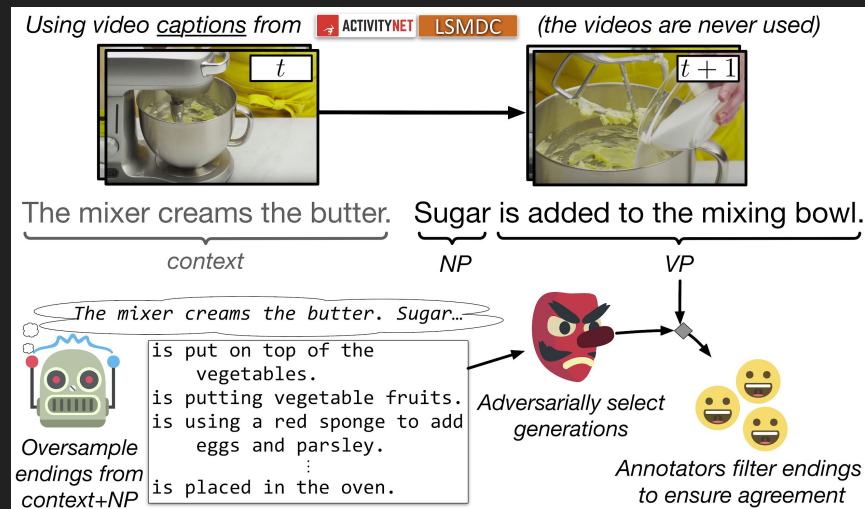
The woman is now blow drying the dog. The dog




- a) is placed in the kennel next to a woman's feet.**
- b) washes her face with the shampoo.
- c) walks into frame and walks towards the dog.
- d) tried to cut her face, so she is trying to do something very close to her face.

SWAG AF =

Situations With Adversarial Generations using Adversarial Filtering

- Collection captions from videos and truncate them after the subject of the second sentence
- Massively oversample a diverse set of potential sentence continuations
- Train a classifier to predict whether a continuation is generated or real.
- To get the “negative” continuations:
 - filter the continuations to the ones the classifier labels as real but humans label as unlikely.
- To get the “positive” continuations
 - filter the continuations to ones humans label as likely.



Rank 	Submission	Created 	Accuracy 
1	RoBERTa <i>Facebook AI</i>	07/18/2019	0.90
2	BigBird <i>Pengcheng He, Weizhu Chen fro...</i>	05/16/2019	0.87
3	BERT (Bidirectional Encoder R... <i>Jacob Devlin, Ming-Wei Chang,...</i>	10/11/2018	0.86
4	BERT-Large-Cased	12/30/2019	0.84
5	OpenAI Transformer Language M... <i>Original work by Alec Radford...</i>	10/11/2018	0.78
6	ESIM with ELMo <i>Zellers, Rowan and Bisk, Yona...</i>	08/30/2018	0.59
7	ESIM with Glove <i>Zellers, Rowan and Bisk, Yona...</i>	08/29/2018	0.52
8	Swag: Decomposable Attention ... <i>Zellers, Rowan and Bisk, Yona...</i>	09/06/2018	0.48
9	Unsupervised QA with PT <i>Sathya Aakur from OSU</i>	01/06/2020	0.44
10	Abductive Reasoning for Unsup... <i>Sathyanarayanan Aakur from OS...</i>	01/03/2020	0.41
11	Unsupervised QA with PT -- No... <i>CVPR Group from USF Tampa</i>	04/25/2019	0.38
12	Zero Training, Total Prior: P... <i>CVPR group - USF, Tampa</i>	04/26/2019	0.38
13	Longest Answer <i>jonathanb@allenai.org</i>	08/30/2018	0.24

Homework to be released tonight (or tomorrow morning)

Implement a few approaches on ROCStories.

Explore a BERT model finetuned on SWAG to better understand the dataset's limitations.

Start thinking about what you want to do for the course project.

Sources

Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches <<https://arxiv.org/pdf/1904.01172.pdf>>

Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence <<https://cs.nyu.edu/davise/papers/CommonsenseFinal.pdf>>

Swag: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference <<https://arxiv.org/pdf/1808.05326.pdf>>

Story Cloze Test and ROCStories Corpora <<https://cs.rochester.edu/nlp/rocstories/>>