Tackling the Biases in the Story Cloze Test Endings

Rishi Sharma, James F. Allen, Omid Bakhshandeh, Nasrin Mostafazadeh

July 15, 2018

Story Understanding and Story Generation

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

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

Requires commonsense reasoning, going beyond pattern recognition and explicit information extraction.

ROC Stories (Mostafazadeh et al., 2016)

A collection of high quality short five sentence stories. Each story:

- Is realistic
- Has a specific beginning and ending, where something happens in between
- Has nothing irrelevant or redundant to the core story

Story Title	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

Story Cloze Test (Mostafazadeh et al., 2016)

The current benchmark for evaluating story understanding and narrative structure learning.

Story Cloze Task: Given a context of four sentences, predict the ending of the story, i.e. Select from the 'right' and 'wrong' ending choices.

Context	Right Ending	Wrong Ending
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 devised a plan for repayment.	Jim decided to open another credit card.

From now on we will refer to SCT as SCT-v1.0

Results On SCT-1.0

Baseline Results

Constant Choose First	0.513
Frequency	0.520
N-gram-overlap	0.494
GenSim	0.539
Sentiment-Full	0.492
Sentiment-Last	0.522
Skip-thoughts	0.552
Narrative-Chains-AP	0.478
Narrative-Chains-Stories	0.494
DSSM	0.585
Human	1.0

LSDSem'17 and Other Models

cogcomp	Logistic	0.776
msap	Logistic	0.752
tbmihaylov	LSTM	0.728
ukp	BiLSTM	0.717
acoli	SVM	0.700
roemmele	RNN	0.672
mflor	Rule Based	0.621
Pranav Goel	Logistic	0.604

cogcomp(UIUC) - Linear classiftcation system that measures a story's coherence based on the sequence of events, emotional trajectory, and plot consistency (includes endings).

msap(UW) - Linear classifier based on language modeling probabilities of the entire story, and linguistic features of only the ending sentences.

Story Ending Biases

Mostafazadeh et al. (2016) were very careful with the task design, the data collection process, and establishing various baselines

- sampled from ROC Stories
- created Wrong Ending stories through Amazon MTurk
- had an AMT to verify quality

Despite that, Schwartz et al. found stylistic differences between right and wrong endings:

- number of words
- n-gram distribution
- character n-gram distribution

Their classifier without feeding context achieves **72.4% accuracy on SCT-v1.0!**

***similar results confirmed by other models, (Cai et al., 2017)*

Biases in Various AI Datasets

From NLI, to VQA, and now Story Cloze Test, our narrow benchmarks inevitably have data creation artifacts and hence yield biased models.

Our Main Contributions

The summary of this talk

1. Analyzed SCT-v1.0 ending features
2. Developed a strong classifier on SCT-v1.0 using only ending features
3. Developed a new crowd-sourcing scheme to tackle the ending biases
4. Collected a new dataset, SCT-v1.5

Statistical Analysis of Endings

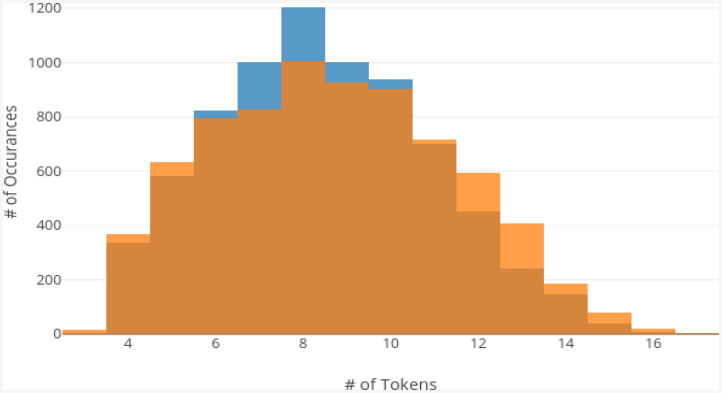
We did an extensive analysis comparing the Right Endings and Wrong Ending features:

- ❏ Token count
- ❏ Sentiment
- ❏ Complexity
- ❏ Token n-grams
- ❏ Character n-grams
- ❏ Part of Speech n-grams
- ❏ Combined Token + POS n-grams

Analysis was done by performing

- A **two sample t-test** between token count, sentiment, an complexity
- **Count measurements** for the n-grams between Right and Wrong Endings

Analysis: Token Count



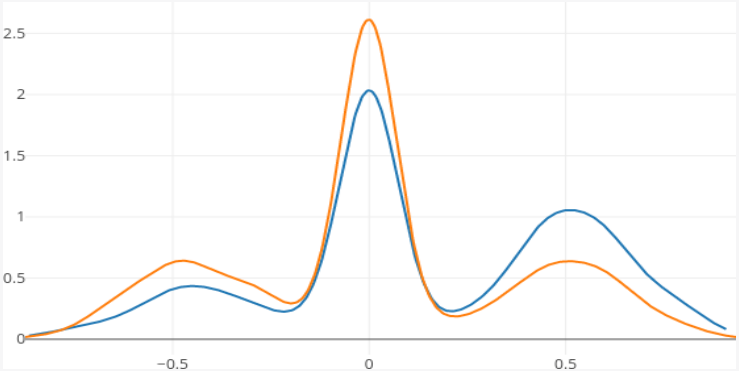
	right endings	wrong endings
token count	8.705	8.466

p-value = 6.63×10^{-5}

Conclusion: Right Endings tend to be longer than Wrong Endings.

Analysis: Sentiment Analysis

Used the Stanford Sentiment Analyzer [0-4] and Vader Sentiment Tagger [-1,1].

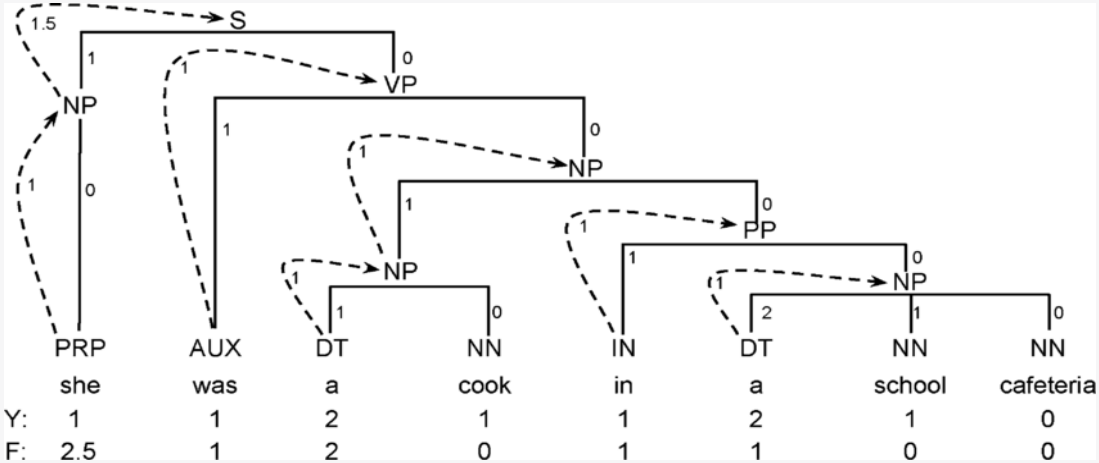


	right endings	wrong endings	p-value
Stanford	2.04	2.02	0.526
VADER	0.146	0.011	3.48×10^{-54}

Conclusion:

- VADER Sentiment score is significant, right endings tend to be more positive than wrong endings.
- The of most stories would probably yield neutral to positive higher and more concentrated peak around Right Endings wider distribution of Right Endings

Analysis: Syntactic Complexity Measurement



	right endings	wrong endings	p-value
Frazier	1.09	1.08	0.135
Yngve	1.15	1.17	0.089

Conclusion: Yngve score was generally more stable and Wrong Endings are more complex than Right Endings.

Image from Roark et al. 2014

Token n-grams

1-5 length stemmed token n-grams, with START token

Character n-grams

4 character size n-grams

Part of Speech n-grams

POS tag and bucketed

Combined Token + POS n-grams

Analysis:

- “got” or “learn” often in Right “decid” often in Wrong
- Wrong frequently have tokens like “n’t ” or “sn’t”
- Right Endings are more likely to feature pronouns (PRP) whereas Wrong Endings are likely to use the proper noun (NNP).

EndingReg Model

A Logistic Regression Model to perform the Story Cloze Test using only the following features extracted from the endings:

- Number of tokens
- VADER sentiment score
- Yngve complexity score
- Token-POS n-grams
- POS n-grams
- Four length character-grams

**also added an L2 regularization penalty and used a grid search was conducted for parameter tuning*

Results on SCT-v1.0

	tokencount, VADER, yngve	ngram	pos	chargrams	All
accuracy	50.3%	69.7%	68.7%	63.4%	71.5%

***the above results indicate the minimum classification accuracy after 10 runs of the EndingReg Model build and classification*

The Criteria for the New Dataset

The Right and Wrong Endings should:

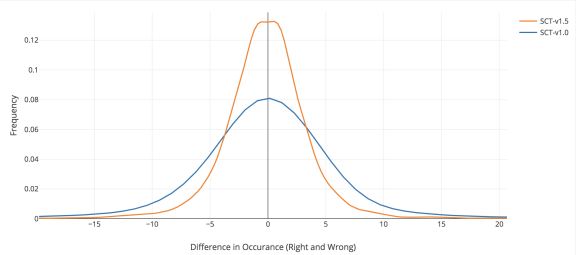
- Contain a similar number of tokens
- Have similar distributions of token n-grams and char-grams
- Occur as standalone events with the same likelihood to occur, with topical, sentimental, or emotional consistencies when applicable.

Collecting The New Dataset

After various rounds of pilot studies, we found the following paradigm to work the best:
New Data Collection Steps:

1. collected 5,000 new five sentence stories with MTurk
2. second AMT round to modify the last sentence to make non-sensible story. Here, the prompt instructs the workers to make sure:
   a. Wrong Ending makes sense standalone
   b. the Right and Wrong ending do not differ in # of words by >3
   c. changes cannot be as simple as negating the verb
3. third AMT to verify quality

This entire process resulted in creating the Story Cloze Test v1.5 (SCT-v1.5) dataset, consisting of 1,571 stories for each validation and test sets.



	token + POS n-gram	char-gram	POS n-gram
SCT-v1.0	13.9	12.4	16.4
SCT-v1.5	7.0	6.3	7.5

Standard deviation of the word and character n-gram counts, as well as the part of speech (POS) counts, between the right and wrong endings.

EndingReg Results

	SCT-v1.0 Val	SCT-v1.0 Test	SCT-v1.5 Test
cogcomp	0.751	0.776	0.608
EndingReg	N/A	0.715	**0.558**
msap	N/A	0.724	0.556
Human	1.0	1.0	1.0

Classification accuracy for various models on the SCT-v1.0 and SCT-v1.5 datasets.

The SOTA Models

Results				
#	User	Entries	Date of Last Entry	PercentageScore ▲
1	jose77	3	07/01/18	0.866382 (1)
2	Hustle	2	06/07/18	0.850267 (2)
3	cogcomp	27	04/15/17	0.776056 (3)
4	msap	8	02/02/17	0.752004 (4)

In *Improving Language Understanding by Generative Pre-Training* model achieves accuracy of 86.5 on SCT-v1.0!

- ➢ Pretrained language model made with Transformer network
- ➢ Task specific supervised learning approach to classify

Initial results on SCT-1.5 show an accuracy of 81.06% for this model, which suggests a deeper story understanding model that goes beyond leveraging the intricacies of the particular test sets.

in Radford, Alec, et al. "Improving Language Understanding by Generative Pre-Training."

Conclusion

In summary

- We presented a comprehensive analysis of the stylistic features isolated in the endings of the original Story Cloze Test (SCT-v1.0).
- Developed a strong classifier using only the story endings
- Developed a new data collection schemes for tackling the stylistic ending features
- Created a new SCT dataset, SCT-v1.5, which overcomes some of the biases.

Takeaways:

- The success of our modified data collection method shows how extreme care must be given for sourcing new datasets.
- However, as shown in multiple AI tasks, no collected dataset is entirely without its inherent biases and often the biases in datasets go undiscovered.

Remember:

- There is still a wide gap between system and human performance, on either SCT 1.0 or SCT 1.5 ;)

Next Steps

- We believe that evaluation benchmarks should evolve and improve over time and we are planning to incrementally update the Story Cloze Test benchmark.
  - Stay tuned for updates on the dataset and SOTA models via http://cs.rochester.edu/nlp/rocstories/
- We expect to release the final dataset, along with reporting the performance of the most recent SCT 1.0 SOTA models on the new dataset, shortly after ACL

Thanks for your attention!
Any Questions?