Tackling the Story Ending Biases in The Story Cloze Test

Rishi Sharma[1], **James F. Allen**[1,2], **Omid Bakhshandeh**[3], **Nasrin Mostafazadeh**[4*]
1 University of Rochester, 2 Institute for Human and Machine Cognition, 3 Verneek.ai 4 Elemental Cognition
rishi.sharma@rochester.edu, nasrinm@cs.rochester.edu

Abstract

The Story Cloze Test (SCT) is a recent framework for evaluating story comprehension and script learning. There have been a variety of models tackling the SCT so far. Although the original goal behind the SCT was to require systems to perform deep language understanding and commonsense reasoning for successful narrative understanding, some recent models could perform significantly better than the initial baselines by leveraging human-authorship biases discovered in the SCT dataset. In order to shed some light on this issue, we have performed various data analysis and analyzed a variety of top performing models presented for this task. Given the statistics we have aggregated, we have designed a new crowdsourcing scheme that creates a new SCT dataset, which overcomes some of the biases. We benchmark a few models on the new dataset and show that the top-performing model on the original SCT dataset fails to keep up its performance. Our findings further signify the importance of benchmarking NLP systems on various evolving test sets.

1 Introduction

Story comprehension has been one of the longest-running ambitions in artificial intelligence (Dijk, 1980; Charniak, 1972). One of the challenges in expanding the field had been the lack of a solid evaluation framework and datasets on which comprehension models can be trained and tested. Mostafazadeh et al. (2016) introduced the Story Cloze Test (SCT) evaluation framework to address this issue. This test evaluates a story comprehension system where the system is given a four-sentence short story as the 'context' and two alternative endings and to the story, labeled 'right ending' and 'wrong ending.' Then, the system's task is to choose the right ending. In order to support this task, Mostafazadeh et al. also provide the ROC Stories dataset, which is a collection of crowd-sourced complete five sentence stories through Amazon Mechanical Turk (MTurk). Each story follows a character through a fairly simple series of events to a conclusion.

Several shallow and neural models, including the state-of-the-art script learning approaches, were presented as baselines (Mostafazadeh et al., 2016) for tackling the task, where they show that all their models perform only slightly better than a random baseline suggesting that richer models are required for tackling this task. A variety of new systems were proposed (Mihaylov and Frank, 2017; Schenk and Chiarcos, 2017; Schwartz et al., 2017b; Roemmele et al., 2017) as a part of the first shared task on SCT at LSDSem'17 workshop (Mostafazadeh et al., 2017). Surprisingly, one of the models made a staggering improvement of 15% to the accuracy, partially due to using stylistic features isolated in the ending choices (Schwartz et al., 2017b), discarding the narrative context. Clearly, this success does not seem to reflect the intent of the original task, where the systems should leverage narrative understanding as opposed to the statistical biases in the data. In this paper, we study the effect of such biases between the ending choices and present a new scheme to reduce such stylistic artifacts.

The contribution of this paper is threefold: (1) we provide an extensive analysis of the SCT dataset to shed some light on the ending data characteristics (Section 3) (2) we develop a new strong classifier for tackling the SCT that uses a variety

* This work was performed at University of Rochester.

Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 752–757
Melbourne, Australia, July 15 - 20, 2018.

Context	Right Ending	Wrong Ending
Ramona was very unhappy in her job. She asked for a raise, but was denied. The refusal prompted her to aggressively comb the want ads. She found an interesting new possibility and set up an interview.	She was offered the new job at a higher salary.	Ramona had no reason to want to change jobs anymore.
The teacher was walking with a stack of papers. Outside started to rain. When the teacher tried to walk down a few steps, she ended up falling. The papers flew out of her hands and landed on the ground.	A passer-by helped her up and helped her collect the papers.	The teacher got up and walked home leaving the papers behind.

Table 1: Example Story Cloze Test examples from the SCT-v1.0 corpus.

of features inspired by all the top-performing systems on the task (Section 4) (3) we design a new crowd-sourcing scheme that yields a new SCT dataset; we benchmark various models on the new dataset (Section 5). The results show that the top-performing SCT system on the the leaderboard[1] (Chaturvedi et al., 2017) fails to keep up the performance on our new dataset. We discuss the implications of this experiment to the greater research community in terms of data collection and benchmarking practices in Section 6. All the code and datasets for this paper will be released to the public. We hope that the availability of the new evaluation set can further support the continued research on story understanding.

2 Related Work

This paper mainly extends the work on creating the Story Cloze Test set (Mostafazadeh et al., 2016), hereinafter SCT-v1.0. The SCT-v1.0 dataset was created as follows: full five-sentence stories from the ROC Stories corpus were sampled, then, the initial four sentences were shown to a set of MTurk[2] crowd workers who were prompted to author 'right' and 'wrong' endings. Mostafazadeh et al. (Mostafazadeh et al., 2016) give special care to make sure there were no boundary cases for 'right' and 'wrong' endings by implementing extra rounds of data filtering. The resulting SCT-v1.0 dataset had a validation (hereinafter, SCT-v1.0 Val) and a test set (SCT-v1.0 test), each with 1,871 cases. Table 1 shows two example story cloze test cases from SCT-v1.0 corpus. As for positive training data, they had provided a collection of 100K five sentence stories. Human performance is reported to be 100% on SCT-v1.0.

Mostafazadeh et al. (2016) provide a variety of baseline models for SCT-v1.0, with the best model performing with an accuracy of 59%. The first shared task on SCT-v1.0 was conducted at the LSDSem'17 workshop (Mostafazadeh et al., 2017), where most of the models performed with 60-70% accuracy. One of the top-performing models, *msap* (Schwartz et al., 2017b,a), built a classifier using linguistic features that have been previously useful in authorship style detection, using only the ending sentences. They used stylistic features such as sentence length, word, and character level n-grams for each ending (fully discarding the context), achieving an accuracy of 72%. In conjunction with their work, Cai et al., (Cai et al., 2017) reported similar observations separately, exposing that features such as sentiment, negation, and length are different between the right and wrong endings. The best model on SCT-v1.0 to this date is *cogcomp*, which is a linear model that uses event sequences, sentiment trajectory, and topical consistency as features, and performs with an accuracy of 77.6%.

This paper takes all their analysis further and introduces a model aggregating all the pinpointed features to shed more light into the stylistic biases isolated in SCT-v1.0 endings.

3 Stylistic Feature Analysis

Despite all the efforts made in the original SCT paper, there was never an extensive analysis of the features isolated in the endings of the stories. We explored the differences among stylistic features such as word-token count, sentiment, and the sentence complexity between the endings, to determine a composite score for identifying sources of bias. For determining the sentiment, we used Stanford CoreNLP (Manning et al., 2014) and the VADER sentiment analyzer (Hutto and Gilbert, 2014). For measuring the syntactic complexity, we used Yngve and Frazier metrics (Yngve, 1960; Frazier, 1985). Table 2 compares these statistics between the right and wrong endings in the SCT-v1.0 dataset. The feature distribution plots can be found in the supplementary material.

[1] As of 15th February 2018.

[2] http://mturk.com

	# of Tokens	Stanford Sentiment	VADER Sentiment	Frazier	Yngve
Right ending	8.705	2.04	0.146	1.09	1.15
Wrong ending	8.466	2.02	0.011	1.08	1.17
p-value	6.63×10^{-5}	0.526	3.48×10^{-54}	0.135	0.089

Table 2: The mean value for the 'right endings' and the 'wrong endings' for the two sample T-tests conducted for each feature.

Furthermore, we conducted an extensive n-gram analysis, using word tokens, characters, part-of-speech, and token-POS (similar to Schwartz et al. (Schwartz et al., 2017b)) as features. We see char-grams such as "sn't" and "not" appear more commonly in the 'wrong endings', suggesting heavy negation. In 'right endings', pronouns are used more frequently versus proper nouns used in 'wrong endings'. Artifacts such as 'pizza' are common in 'wrong endings,' which could suggest that for a given topic, the authors may replace an object in a right ending with a wrong one and quickly think up a common item such as pizza to create a 'wrong' one. An extensive analysis of these features, including the n-gram analysis, can be found in the supplementary material.

4 Model

Following the analysis above, we developed a Story Cloze model, hereinafter EndingReg, that only uses the ending features while disregarding the story context for choosing the right ending. We expanded each Story Cloze Test case's ending options into a set of two single sentences. Then, for each sentence, we created the following features:

1. Number of tokens
2. VADER composite sentiment score
3. Yngve complexity score
4. Token-POS n-grams
5. POS n-grams
6. Four length character-grams

All n-gram features needed to appear at least five times throughout the dataset. The features were collected for each five-sentence story and then fed into a logistic regression classifier. As an initial experiment, we trained this model using the SCT-v1.0 validation set and tested on the SCT-v1.0 test set. An L2 regularization penalty was used to enforce a Gaussian prior on the feature-space, where a grid search was conducted for hyper-parameter tuning. This model achieves an accuracy of 71.5% on the SCT-v1.0 dataset which is on par with the highest score achieved by any model using only the endings. Table 3 shows the accuracies obtained by models using only those particular features. We achieve minimal but sometimes important classification using token count, VADER, and Yngve in combination alone, better classification using POS or char-grams alone, and best classification using n-grams alone. By combining all of them we achieve the overall best results.

token-count+VADER+yngve	ngram	pos	char-grams	All
50.3%	69.7%	68.7%	63.4%	71.5%

Table 3: Classification results on SCT-v1.0 using each of the feature sets designated in the columns.

5 Data Collection

Based on the findings above, a new test set for the SCT was deemed necessary. The premise of predicting an ending to a short story, as opposed to predicting say a middle sentence, enables a more systematic evaluation where human can agree on the cases 100%. Hence, our goal was to come up with a data collection scheme that overcomes the data collection biases, while keeping the original evaluation format. As the data analysis revealed, the token count, sentiment, and the complexity are not as important features for classification as the ending n-grams are. We set the following goals for sourcing the new 'right' and 'wrong' endings. They both should:

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

First, we crowdsourced 5,000 new five-sentence stories through Amazon Mechanical Turk. We prompted the users in the same manner described in Mostafazadeh et al. (2016). In order to source new 'wrong' endings, we tried two different methods. In Method #1, we kept the original ending sourcing format of Mostafazadeh et al., but imposed some further restrictions. This was done

by taking the first four sentences of the newly collected stories and asking an MTurker to write a 'right' and 'wrong' ending for each. The new restrictions were: 'Each sentence should stay within the same subject area of the story,' and 'The number of words in the Right and Wrong sentences should not differ by more than 2 words,' and 'When possible, the Right and Wrong sentences should try to keep a similar tone/sentiment as one another.' The motivation behind this technique was to reduce the statistical differences by asking the user to be mindful of considerations.

In Method #2, we took the five sentences stories and prompted a second set of MTurk workers to modify the fifth sentence in order to make a resulting five-sentence story non-sensible. Here, the prompt instructs the workers to make sure the new 'wrong ending' sentence makes sense standalone, that it does not differ in the number of words from the original sentence by more than three words, and that the changes cannot be as simple as e.g., putting the word 'not' in front of a description or a verb. As a result, the workers had much less flexibility for changing the underlying linguistic structures which can help tackle the authorship style differences between the 'right' and 'wrong' endings.

The results in Table 4, which show classification accuracy when using EndingReg on the two new data sources, show that Method #2 is a slightly better data sourcing scheme in reducing the bias, since the EndingReg model's performance is slightly worse. The set was further filtered through human verification similar to Mostafazadeh et al. (2016). The filtering was done by splitting each SCT-v1.0's two alternative endings into two independent five-sentence stories and asking three different MTurk users to categorize the story as either: one where the story made complete sense, one where the story made sense until the last sentence and one where the story does not make sense for another reason. Stories were only selected if all the three MTurk users verified that the story with the 'right ending' and the corresponding story with the 'wrong ending' were verified to be indeed right and wrong respectively. This ensured a higher quality of data and eliminating boundary cases. 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.

	Method #1	Method #2
EndingReg	0.709	0.695
cogcomp	0.649	0.641

Table 4: Comparison of initial data sourcing methods

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

Table 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.

6 Results

In order to test the decrease in n-gram bias, which was the most salient feature for the classification task using only the endings, we compare the variance between the n-gram counts from SCT-v1.0 to SCT-v1.5. The results are presented in Table 5, which indicates the drop in the standard deviations in our new dataset. Table 6 shows the classification results of various models on SCT-v1.5. The drop in accuracy of the EndingReg model between the SCT-v1.0 and SCT-v1.5 shows a significant improvement on the statistical weight of the stylistic features generated by the model.

Since the main features used are the token length and the various n-grams, this suggests that the new 'right endings' and 'wrong endings' have much more similar token n-gram, pos n-gram, pos-token n-gram and char-gram overlap. Furthermore, the CogComp model's performance has significantly dropped on SCT-v1.5. Although this model seems to be using story comprehension features such as event sequencing, since the endings are included in the sequences, the biases within the endings have influenced the predictions and the weak performance of the model in SCT-v1.5 suggest that this model had picked up on the biases of SCT-v1.0 as opposed to really understanding the context. In particular, the posterior probabilities for each ending choice using their features are quite similar on the SCT-v1.5. These results place the classification accuracy of this top performing model on par with or worse than the models that did not use the ending features of the old SCT-v1.0 dataset (Mostafazadeh et al., 2017), which suggest that the gap that once was held by models using the ending biases seems to be corrected for. Al-

	SCT-v1.0 Val	SCT-v1.0 Test	SCT-v1.5 Test
cogcomp	0.751	0.776	0.608
EndingReg	N/A	0.715	**0.644**
average sentiment	0.489	0.492	0.496
last sentiment	0.514	0.522	0.525
word2vec	0.545	0.539	0.594
human	1.0	1.0	1.0

Table 6: Classification accuracy for various models on the SCT-v1.0 and SCT-v1.5 datasets.

though we did not get to test all the other models published on SCT-v1.0 directly, we predict similar trends.

It is important to point out that the 64.4% performance attained by our EndingReg model is still high for a model which completely discards the context. This indicates that although we could correct for some of the stylistic biases, there are some other hidden patterns in the new endings that would not have been accounted for without having the EndingReg baseline. This showcases the importance of maintaining benchmarks that evolve and improve over time, where systems should not be optimized for particular narrow test sets. We propose the community to report accuracies on both SCT-v1.0 and SCT-v1.5, both of which still have a huge gap between the best system and the human performance.

7 Conclusion

In this paper, we presented a comprehensive analysis of the stylistic features isolated in the endings of the original Story Cloze Test (SCT-v1.0). Using that analysis, along with a classifier we developed for testing new data collection schemes, we created a new SCT dataset, SCT-v1.5, which overcomes some of the biases. Based on the results presented in this paper, we believe that our SCT-v1.5 is a better benchmark for story comprehension. However, as shown in multiple AI tasks (Ettinger et al., 2017; Antol et al., 2015; Jabri et al., 2016; Poliak et al., 2018), no collected dataset is entirely without its inherent biases and often the biases in datasets go undiscovered. We believe that evaluation benchmarks should evolve and improve over time and we are planning to incrementally update the Story Cloze Test benchmark. All the new versions, along with a leader-board showcasing the state-of-the-art results, will be tracked via CodaLab https://competitions.codalab.org/competitions/15333.

The success of our modified data collection method shows how extreme care must be given for sourcing new datasets. We suggest the next SCT challenges to be completely blind, where the participants cannot deliberately leverage any particular data biases. Along with this paper, we are releasing the datasets and the developed models to the community. All the announcements, new supplementary material, and datasets can be accessed through http://cs.rochester.edu/nlp/rocstories/. We hope that this work ignites further interest in the community for making progress on story understanding.

Acknowledgement

We would like to thank Roy Schwartz for his valuable feedback regarding some of the experiments. We also thank the amazing crowd workers, without the work of whom this work would have been impossible. This work was supported in part by grant W911NF15-1-0542 with the US Defense Advanced Research Projects Agency (DARPA) as a part of the Communicating with Computers (CwC) program.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Zheng Cai, Lifu Tu, and Kevin Gimpel. 2017. Pay attention to the ending: Strong neural baselines for the ROC story cloze task. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Eugene Charniak. 1972. Toward a model of children's story comprehension.

Snigdha Chaturvedi, Haoruo Peng, and Dan Roth. 2017. Story comprehension for predicting what happens next.

Teun A. Van Dijk. 1980. Story comprehension: An introduction. *Poetics*, 9(1):1 – 21. Special Issue Story Comprehension.

Allyson Ettinger, Sudha Rao, Hal Daum III, and Emily M Bender. 2017. Towards linguistically generalizable nlp systems: A workshop and shared task.

In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10.

Lyn Frazier. 1985. Natural language parsing. *Cambridge University Press*.

C.J. Hutto and E.E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media*.

Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *ECCV*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. Stanford corenlp natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Todor Mihaylov and Anette Frank. 2017. Simple story ending selection baselines. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, Valencia, Spain. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL HLT 2016*, page 839849.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James F. Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis Only Baselines in Natural Language Inference. In *Joint Conference on Lexical and Computational Semantics (StarSem)*.

Melissa Roemmele, Sosuke Kobayashi, Naoya Inoue, and Andrew Gordon. 2017. An RNN-based binary classifier for the story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, Valencia, Spain. Association for Computational Linguistics.

Niko Schenk and Christian Chiarcos. 2017. Resource-lean modeling of coherence in commonsense stories. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, Valencia, Spain. Association for Computational Linguistics.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A. Smith. 2017a. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proc. of CoNLL*.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A. Smith. 2017b. Story cloze task: Uw nlp system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*.

Victor Yngve. 1960. A model and an hypothesis for language structure.