LSDSem 2017 Shared Task: The Story Cloze Test

Nasrin Mostafazadeh[1], **Michael Roth**[2,3], **Annie Louis**[4],
Nathanael Chambers[5], **James F. Allen**[1,6]

1 University of Rochester 2 University of Illinois at Urbana-Champaign 3 University of Edinburgh
4 University of Essex 5 United States Naval Academy 6 Florida Institute for Human & Machine Cognition
{nasrinm,james}@cs.rochester.edu mroth@coli.uni-saarland.de
aplouis@essex.ac.uk nchamber@usna.edu

Abstract

The LSDSem'17 shared task is the Story Cloze Test, a new evaluation for story understanding and script learning. This test provides a system with a four-sentence story and two possible endings, and the system must choose the correct ending to the story. Successful narrative understanding (getting closer to human performance of 100%) requires systems to link various levels of semantics to commonsense knowledge. A total of eight systems participated in the shared task, with a variety of approaches including end-to-end neural networks, feature-based regression models, and rule-based methods. The highest performing system achieves an accuracy of 75.2%, a substantial improvement over the previous state-of-the-art.

1 Introduction

Building systems that can understand stories or can compose meaningful stories has been a long-standing ambition in natural language understanding (Charniak, 1972; Winograd, 1972; Turner, 1994; Schubert and Hwang, 2000). Perhaps the biggest challenge of story understanding is having commonsense knowledge for comprehending the underlying narrative structure. However, rich semantic modeling of the text's content involving words, sentences, and even discourse is crucially important. The workshop on Linking Lexical, Sentential and Discourse-level Semantics (LSDSem)[1] is committed to encouraging computational models and techniques which involve multiple levels of semantics.

[1] http://www.coli.uni-saarland.de/~mroth/LSDSem/

The LSDSem'17 shared task is the Story Cloze Test (SCT; Mostafazadeh et al., 2016). The SCT is one of the recent proposed frameworks on evaluating story comprehension and script learning. In this test, the system reads a four-sentence story along with two alternative endings. It is then tasked with choosing the correct ending. Mostafazadeh et al. (2016) summarize the outcome of experiments conducted using several models including the state-of-the-art script learning approaches. They suggest that current methods are only slightly better than random performance and more powerful models will require richer modeling of the semantic space of stories.

Given the wide gap between human (100%) and state-of-the-art system (58.5%) performance, the time was ripe to hold the first shared task on SCT. In this paper, we present a summary on the first organized shared task on SCT with eight participating systems. The submitted approaches to this non-blind challenge ranged from simple rule-based methods, to linear classifiers and end-to-end neural models, to hybrid models that leverage a variety of features on different levels of linguistic analysis. The highest performing system achieves an accuracy of 75.2%, which substantially improves the previously established state-of-the-art. We hope that our findings and discussions can help reshape upcoming evaluations and shared tasks involving story understanding.

2 The Story Close Test (SCT)

In the SCT task, the system should choose the right ending to a given four-sentence story. Hence, this task can be seen as a reading comprehension test in which the binary choice question is always, 'Which of the two endings is the most plausible correct ending to the story?'. Table 1 shows three example SCT cases.

Context	Right Ending	Wrong Ending
Sammy's coffee grinder was broken. He needed something to crush up his coffee beans. He put his coffee beans in a plastic bag. He tried crushing them with a hammer.	It worked for Sammy.	Sammy was not that much into coffee.
Gina misplaced her phone at her grandparents. It wasnt anywhere in the living room. She realized she was in the car before. She grabbed her dads keys and ran outside.	She found her phone in the car.	She didnt want her phone anymore.
Sarah had been dreaming of visiting Europe for years. She had finally saved enough for the trip. She landed in Spain and traveled east across the continent. She didn't like how different everything was.	Sarah decided that she preferred her home over Europe.	Sarah then decided to move to Europe.

Table 1: Example Story Cloze Test instances from the Spring 2016 release.

Story Title	Story
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.
Marco Votes For President	Marco was excited to be a registered voter. He thought long and hard about who to vote for. Finally he had decided on his favorite candidate. He placed his vote for that candidate. Marco was proud that he had finally voted.
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.

Table 2: Example ROCStories instances from the Winter 2017 release.

As described in Mostafazadeh et al. (2016), the SCT cases are collected through Amazon Mechanical Turk (Mturk) on the basis of the ROCStories corpus, a collection of five-sentence everyday life stories which are full of stereotypical sequence of events. To construct SCT cases, they randomly sampled complete five-sentence stories from the ROCStories corpus and presented only the first four sentences of each story to the Mturk workers. Then, for each story, a worker was asked to write a 'right ending' and a 'wrong ending'. This resulting set was further filtered by human verification: they compile each SCT case into two independent five-sentence stories, once with the 'right ending' and once with the 'wrong ending'. Then, for each story they asked three crowd workers to verify if the given five-sentence story makes sense as a meaningful story, rating on the scale of {-1, 0, 1}. Then they retain the cases in which the 'right ending' had three 1 ratings and the 'wrong ending' had three 0 ratings. This verification step ensures that there are no boundary cases of 'right ending' and 'wrong ending' for human. Finally, any stories used in creating this SCT set are removed from the original ROCStories corpus.

3 Shared Task Setup

For the shared task, we provided the same dataset as created by Mostafazadeh et al. (2016), which consists of a development and test set each containing 1,871 stories with two alternative endings. At this stage, we used this already existing non-blind dataset with established baselines to build up momentum for researching the task. This dataset can be accessed through `http://cs.rochester.edu/nlp/rocstories/`.

As the training data, we released an extended set of ROCStories[2], called ROCStories Winter 2017. We followed the same crowdsourcing setup described in Mostafazadeh et al. Table 2 provides three example stories in this dataset. As these examples show, these are complete stories and do not come with a wrong ending[3]. Although we provided the additional ROCStories, the participants were encouraged to use or construct any training data of their choice. Overall, the participants were provided three datasets with the statistics listed in Table 3.

Following Mostafazadeh et al. (2016), we eval-

[2]The extended ROCStories dataset can be accessed via `http://cs.rochester.edu/nlp/rocstories/`.

[3]The ROCStories corpus can be used for a variety of applications ranging from story generation to script learning.

ROCStories (training data)	**98,159**
Story Cloze validation set, Spring 2016	**1,871**
Story Cloze test set, Spring 2016	**1,871**

Table 3: The size of the provided shared task datasets.

uate the systems in terms of accuracy, which we measure as $\frac{\#correct}{\#test\ cases}$. Any other details regarding our shared task can be accessed via our shared task page `http://cs.rochester.edu/nlp/rocstories/LSDSem17/`.

4 Submissions

The Shared Task was conducted through CodaLab competitions[4]. We received a total of 18 registrations, out of which eight teams participated: four teams from the US, three teams from Germany and one team from India.

In the following, we provide short paragraphs summarizing our baseline and approaches of the submissions. More details can be found in the respective system description papers.

msap (University of Washington). Linear classifier based on language modeling probabilities of the entire story, and linguistic features of only the ending sentences (Schwartz et al., 2017). These ending "style" features include sentence length as well as word and character n-gram in each candidate ending (independent of story). These style features have been shown useful in other tasks such as age, gender, or native language detection.

cogcomp (University of Illinois). Linear classification system that measures a story's coherence based on the sequence of events, emotional trajectory, and plot consistency. This model takes into account frame-based and sentiment-based language modeling probabilities as well as a topical consistency score.

acoli (Goethe University Frankfurt am Main) and **tbmihaylov (Heidelberg University).** Two resource-lean approaches that only make use of pretrained word representations and compositions thereof (Schenk and Chiarcos, 2017; Mihaylov and Frank, 2017). Composition functions are learned as part of a feed-forward and LSTM neural networks, respectively.

[4]The Story Cloze CodaLab page can be accessed here: `https://competitions.codalab.org/competitions/15333`

ukp (Technical University of Darmstadt). Combination of a neural network-based (Bi-LSTM) classifier and a traditional feature-rich approach (Bugert et al., 2017). Linguistic features include aspects of sentiment, negation, pronominalization and n-gram overlap between the story and possible endings.

roemmele (University of Southern California). Binary classifier based on a recurrent neural network that operates over (sentence-level) Skip-thought embeddings (Roemmele et al., 2017). For training, different data augmentation methods are explored.

mflor (Educational Testing Service). Rule-based combination of two systems that score possible endings in terms of how well they lexically cohere with and fit the sentiment of the given story (Flor and Somasundaran, 2017). Sentiment is given priority, and the model backs off to lexical coherence based on pointwise mutual information scores.

Pranav_Goel (IIT Varanasi). Ensemble model that takes into account scores from two systems that measure overlap in sentiment and sentence similarity between the story and the two possible endings (Goel and Singh, 2017).

ROCNLP (baseline) Two feed-forward neural networks trained jointly on ROCStories to project the four-sentences context and the right fifth sentence into the same vector space. This model is called Deep Structured Semantic Model (DSSM) (Huang et al., 2013) and had outperformed all the other baselines reported in Mostafazadeh et al. (2016).

5 Results

An overview of the models and the resources used in each participating system, along with their quantative results, is given in Table 4. Given that the DSSM model was previously trained on about 50K ROCStories, we retrained this model on our full dataset of 98,159 stories. We include the results of this model under ROCNLP in Table 4. With accuracy values in a range from 60% to 75.2%, we observe that all teams outperform the baseline model. The best result in this shared task has been achieved by **msap**, the participating team from the University of Washington.

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	**ROCNLP (baseline)**	DSSM	Spring 2016, Winter 2017	–	–	0.595

Table 4: Overview of models and resources used by the participating teams. For each team only their best performing system on the Spring 2016 Test Set is included, as submitted to CodaLab. Please refer to the system description papers for a list of other models. Human is reported to perform at 100%.

6 Discussion

We briefly highlight some observations regarding modeling choices and results.

Embeddings. All but two teams made use of pretrained embeddings for words or sentences. **tbmihaylov** (Mihaylov and Frank, 2017) experimented with various pretrained embeddings in their resource-lean model and found that the choice of embeddings has a considerable impact on model accuracy. Interestingly, the best participating team used no pretrained embeddings at all.

Neural networks. The six highest scoring models all include neural network architectures in one way or another. While the teams ranked 3–6 attempt to utilize hidden layers directly for prediction, the top two teams use the output of neural language models to generate different combinations of features. Further, while the third place team's best model was an LSTM, their logistic regression classifier with Word2Vec-based features achieved similar performance. The combination of different neural features (including non-neural ones) appears to have made the difference in the top system's ablation tests.

Sentiment. Three teams report concurrently that a sentiment model alone can achieve 60–65% accuracy but performance seems to vary dependent on implementation details. This is notable in that the sentiment baseline which chose the ending with a matching sentiment to the context (presented in Mostafazadeh et al. (2016)) did not achieve accuracy above random chance. One difference is that these more successful approaches used sentiment lexicons to score words and sentences, whereas Mostafazadeh et al. used the automatic sentiment classifier in Stanford's CoreNLP. Finally, **mflor** (Flor and Somasundaran, 2017) analyzed the Story Cloze Test Validation (Spring 2016) set and found that 78% of the stories have sentiment bearing words in the first sentences and in at least one possible ending. Evaluating on that subset showed increased performance, further suggesting that sentiment is an important factor in alternate ending prediction.

Stylistic Features on Endings. One of the models proposed by **msap** (Schwartz et al., 2017) ignored the entire story, building features only from the ending sentences. They trained a linear classifier on the right and wrong ending sentences adopting style features that have been shown useful in other tasks such as gender or native language detection. This model achieved remarkably good performance at 72.4%, indicating that there

are characteristics inherent to right/wrong endings independent of story reasoning. It is not clear whether these results generalize to novel story ending predictions, beyond the particular Spring 2016 sets. Whether this model captures an artifact of the test set creation, or it indicates general features about how stories are ended must remain for future investigation.

Negative results. Some papers describe additional experiments with features and methods that are not part of the submitted system, because their inclusion resulted in sub-optimal performance. For example, **Pranav_Goel** (Goel and Singh, 2017) discuss additional similarity measures based on doc2vec sentence representations (Le and Mikolov, 2014); **tbmihaylov** (Mihaylov and Frank, 2017) experiment with ConceptNet Numberbatch embeddings (Speer and Chin, 2016); and **mflor** (Flor and Somasundaran, 2017) showcase results with alternative sentiment dictionaries such as MPQA (Wilson et al., 2005).

7 Conclusions

All participants in the Story Cloze shared task of LSDSem outperformed the previously published best result of 58.5%, and the new state-of-the-art accuracy dramatically increased to 75.2% with the help of a well-designed RNNLM and unique stylistic features on the ending sentences.

One of the main takeaways from the 8 submissions is that the detection of correct ending sentences requires a variety of different reasoners. It appears from both results and post-analysis that sentiment is one factor in correct detection. However, it is also clear that coherence is critical, as the systems with language models all observed increases in prediction accuracy. Beyond these, the best performing system showed that there are stylistic features isolated in the ending sentences, suggesting yet another area of further investigation for the next phases of this task.

As the first shared task on SCT, we decided not to hold a blind challenge. For the future blind challenges, the question is how robust are the presented approaches to novel test cases and how well can they generalize out of the scope of the current evaluation sets. We speculate that the models which use generic language understanding and semantic cohesion criteria rather than relying on certain intricacies of the testing corpora can generalize more successfully, which should be carefully assessed in future.

Although this shared task was successful at setting a new state-of-the-art for SCT, clearly, there is still a long way towards achieving human-level performance of 100% on even the current test set. We are encouraged by the high level of participation in the LSDSem 2017 shared task, and hope the new models and results encourage further research in story understanding. Our findings can help direct the creation of the next SCT datasets towards enforcing deeper story understanding.

Acknowledgment

We would like to thank the wonderful crowd workers whose daily story writing made collecting the additional dataset possible. We thank Alyson Grealish and Ahmed Hassan Awadallah for their help in the quality control of ROCStories corpus. This work was supported in part by grant W911NF-15-1-0542 with the US Defense Advanced Research Projects Agency (DARPA) as a part of the Communicating with Computers (CwC) program, a grant from the Office of Naval Research, Army Research Office (ARO), and a DFG Research Fellowship (RO 4848/1-1).

References

Michael Bugert, Yevgeniy Puzikov, Andreas Rckl, Judith Eckle-Kohler, Teresa Martin, Eugenio Martnez-Cmara, Daniil Sorokin, Maxime Peyrard, and Iryna Gurevych. 2017. LSDSem 2017: Exploring data generation methods 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.

Eugene Charniak. 1972. *Toward a Model of Children's Story Comprehension*. Ph.D. thesis, MIT.

Michael Flor and Swapna Somasundaran. 2017. Sentiment analysis and lexical cohesion for the story cloze task. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, Valencia, Spain. Association for Computational Linguistics.

Pranav Goel and Anil Kumar Singh. 2017. IIT (BHU): System description for LSDSem'17. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, Valencia, Spain. Association for Computational Linguistics.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep

structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, (CIKM '13), pages 2333–2338, New York, NY, USA. ACM.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China.

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*, pages 839–849, San Diego, California, June. Association for Computational Linguistics.

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.

Lenhart K. Schubert and Chung Hee Hwang. 2000. Episodic logic meets little red riding hood: A comprehensive, natural representation for language understanding. In *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. MIT/AAAI Press.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith. 2017. Story cloze task: UW NLP system. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem)*, Valencia, Spain. Association for Computational Linguistics.

Robert Speer and Joshua Chin. 2016. An ensemble method to produce high-quality word embeddings. *arXiv preprint arXiv:1604.01692*.

Scott R. Turner. 1994. The creative process: A computer model of storytelling and creativity. *Hillsdale: Lawrence Erlbaum.*

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, October.

Terry Winograd. 1972. *Understanding Natural Language*. Academic Press, Inc., Orlando, FL, USA.