Story Cloze Evaluator: Vector Space Representation Evaluation by Predicting What Happens Next

Nasrin Mostafazadeh[1], **Lucy Vanderwende**[2], **Wen-tau Yih**[2], **Pushmeet Kohli**[2], **James Allen**[1,3]
1 University of Rochester, 2 Microsoft Research, 3 The Institute for Human & Machine Cognition
{nasrinm, james}@cs.rochester.edu, {lucyv, scottyih, pkohli}@microsoft.com

Abstract

The main intrinsic evaluation for vector space representation has been focused on textual similarity, where the task is to predict how semantically similar two words or sentences are. We propose a novel framework, Story Cloze Evaluator, for evaluating vector representations which goes beyond textual similarity and captures the notion of predicting what should happen next given a context. This evaluation methodology is *simple to run, scalable, reproducible by the community, non-subjective, 100% agreeable by human, and challenging to the state-of-the-art models*, which makes it a promising new framework for further investment of the representation learning community.

1 Introduction

There has been a surge of work in the vector representation research in the past few years. While one could evaluate a given vector representation (embedding) on various down-stream applications, it is time-consuming at both implementation and runtime, which gives rise to focusing on an intrinsic evaluation. The intrinsic evaluation has been mostly focused on textual similarity where the task is to predict how semantically similar two words/sentences are, which is evaluated against the gold human similarity scores.

It has been shown that semantic similarity tasks do not accurately measure the effectiveness of an embedding in the other down-stream tasks (Schnabel et al., 2015; Tsvetkov et al., 2015). Furthermore, human annotation of similarity at sentence-level without any underlying context can be subjective, resulting in lower inter-annotator agreement and hence a less reliable evaluation method. There has not been any standardized intrinsic evaluation for the quality of sentence and document-level vector representations beyond textual similarity[1]. There is therefore a crucial need for new ways of evaluating semantic representations of language which capture other linguistic phenomena.

In this paper we propose a new proxy task, Story Cloze Test, for measuring the quality of vector space representations for generic language understanding and commonsense reasoning. In this task, given a four-sentence story (called the context) and two alternative endings to the story, the system is tasked with choosing the right ending. We propose the following Story Cloze Evaluator modules: (1) Given an embedding of a four-sentence story (the context) and two alternative ending sentences, this module rewards the system if the embedding of the context is closer to the right ending than the wrong ending. (2) Given the embedding for each of the four sentences and each of the two alternatives, this module uses the trajectory of the four vectors to predict the embedding of the fifth sentence. Then the system is rewarded if the predicted vector is closer to the right ending than the wrong ending.

A vector representation that achieves a high score according to the Story Cloze Evaluator is demonstrating some level of language and narrative understanding. We describe the Story Cloze Test in Section 2, where we show that this test is scalable, non-subjective and 100% agreeable by human. We further describe our evaluation methodology in Section 3. As with any evaluation framework, we expect the setup to be modified over time, the updates of which can be followed through `http://`

[1]Examples of this include the semantic relatedness (SICK) dataset (Marelli et al., 2014), where given two sentences, the task is to produce a score of how semantically related these sentences are

Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, pages 24–29,
Berlin, Germany, August 12, 2016.

cs.rochester.edu/nlp/rocstories/RepEvalPredictingTheNext/.

2 Story Cloze Test: Predicting What Happens Next

Representation and learning of commonsense knowledge is one of the foundational problems for enabling deep language understanding. This issue is the most challenging for understanding casual and correlational relationships between events, and predicting what happens next. A recent framework for evaluating story and script[2] understanding (Schank and Abelson, 1977) is the 'Story Cloze Test' (Mostafazadeh et al., 2016), where given two alternative endings to a four-sentence story (the context), a system is tasked with choosing the right ending. Table 1 shows a few example instances of the Story Cloze Test[3].

Although the Story Cloze Test was initially proposed to evaluate story understanding and script learning capabilities of a system, we see it as a perfect fit for intrinsic evaluation of vector space representation at sentence and paragraph level. The Story Cloze Test is unique in requiring a system to demonstrate generic commonsense understanding about stereotypical causal and temporal relations between daily events, making it a unique proxy task for vector space representation at sentence and paragraph level.

Story Cloze Test looks similar to language modeling at sentence level. However, predicting an ending to a story is less subjective and more deterministic than only predicting the next sentence. Experimental evaluation has shown (Mostafazadeh et al., 2016) that huamn performs 100% on this task, which makes it a very reliable test framework. Moreover, evaluation results have shown that a host of state-of-the-art models struggle to achieve a high score on this test [4], which makes the task even more compelling for the representation learning community to focus on.

2.1 Crowdsourcing Story Cloze Test

Story Cloze Test dataset can be easily scaled to hundreds of thousands of instances by crowdsourcing. The crowdsourcing starts from sampling complete five-sentence stories from the ROCStories corpus. This corpus is a collection of ∼50,000 crowdsourced short commonsense everyday stories [5], each of which has the following major characteristics: (1) is realistic and non-fictional, (2) has a clear beginning and ending where something happens in between, (3) does not include anything irrelevant to the core story. These stories are full of stereotypical causal and temporal relations between events, making them a great resource for commonsense reasoning and generic language understanding.

The crowdsourcing process continues as follows: given a complete five-sentence story, the fifth sentence is dropped and only the first four sentences (the context) are shown to the crowd workers. For each context, a worker was asked to write a 'right ending' and a 'wrong ending'. The workers were prompted to write 'wrong ending' which satisfies two conditions: (1) The sentence should follow up the story by sharing at least one of the characters of the story, and (2) The sentence should be entirely realistic and sensible when read in isolation. These conditions make sure that the Story Cloze Test cases are not trivial.

Quality Control. The accuracy of the Story Cloze test set plays a crucial role in propelling the research community towards the right direction. A two-step quality control step makes sure that there are no vague or boundary cases in the test set. First, the initially collected Story Cloze Test cases are compiled into two sets of full five-sentence stories. Then for each five-sentence story, independently, three crowd workers are tasked to verify whether or not the given sequence of five sentences makes sense as a meaningful and coherent story, rating within {-1, 0, 1}. Then, only the initial test cases which get three ratings of 1 for their 'right ending' compilation and three ratings of -1 for their 'wrong ending' compilation are included in the final dataset. This process ensures that there are no boundary case of vague, incoherent, or hard to follow stories, making human performance of 100% accuracy possible.

Data Split. Any collection of Story Cloze Test instances will be split into validation and test sets[6], where the test set will be blind and not accessible by the systems under evaluation. There is cur-

[2]Scripts represent structured knowledge about stereotypical event sequences together with their participants, e.g., {X kills Y, Y dies, X gets detained}.

[3]More examples can be found here: http://cs.rochester.edu/nlp/rocstories/

[4]The best performing system based on Deep Structured Semantic Model (DSSM) (Huang et al., 2013) performs with the accuracy of 58%, where a random baseline achieves 50%.

[5]These stories can be found via http://cs.rochester.edu/nlp/rocstories

[6]We also consider providing a designated training set, however, different models can choose to use any resources for training.

Context	Right Ending	Wrong Ending
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.
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.
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.
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.
When I first moved into my house, I didn't know my neighbors. While mowing one day, I found a kickball in my yard. I felt this was the perfect opportunity to meet my neighbors. I grabbed the ball and went next door to return it.	They were very friendly and appreciated it.	I threw the kickball through their closed window.
Amber had a lot of things to do this Sunday. She made a list of all the places she needed to go. She hurried to get ready. She was worried that she would not have enough time.	Amber was so hurried that she left the list at home.	Amber enjoyed a relaxing two hour brunch.
Tim was entering a baking contest. He decided to make his famous donuts. He made a big batch and entered them into the contest. The judges thought they were delicious.	Tim won the baking contest.	The judges vomited from the taste of the donuts.

Table 1: Example Story Cloze Test instances.

rently 3,744 instances of Story Cloze Test[7] that showcase our desired quality for the larger dataset.

3 Story Cloze Evaluator

There are various ways we can use Story Cloze Test for evaluating an embedding model at paragraph and sentence level. We propose the following alternatives.

3.1 Joint Paragraph and Sentence Level Evaluator

For this evaluator, a system should have two different modules for embedding either an alternative (a sentence) or a context (a paragraph), which ideally should be trained jointly. The evaluator works as follows: given the vector representations of the two alternative endings and the four-sentence context as a whole (Figure 1), it rewards the embedding model if the context's embedding is closer to the right ending embedding than the wrong ending. The closeness can be measured via cosine similarity of the embeddings.

This method evaluates joint paragraph-level and sentence-level vector representations, where all the representations are projected into the same vector space. Representing semantics of a paragraph as a vector is a major unresolved issue in the field, requiring its own detailed discussions.

[7]Accessible through `http://cs.rochester.edu/nlp/rocstories/`.

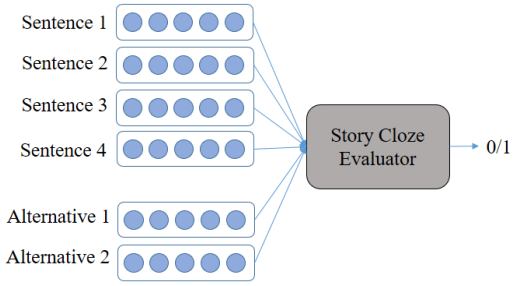

Figure 1: Sentence-level stroy cloze evaluator.

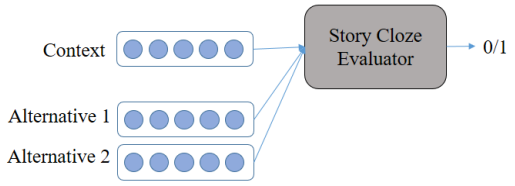

Figure 2: Joint paragraph and sentence level story cloze evaluator.

Here, we represent a paragraph according to what should happen next, which can be beneficial for various generic language comprehension frameworks. Deferring the representation of the context paragraph to the system under evaluation, makes it possible to use various sequence modeling techniques, among others, for representing the context.

3.2 Sentence-level Evaluator

For this evaluator, the embedding should be at sentence-level. The evaluator works as follows: given the vector representations for each of the four sentences and the two alternative endings (Figure 2), the evaluator component uses the trajectory of the four sentences to predict the embedding of the ending sentences. Then the embedding model is rewarded if the predicted embedding is closer to the right ending than the wrong ending.

Given that the evaluator module should be simple and deterministic, we do not want to use any learning components inside the evaluator. Hence, we need a simple and deterministic procedure for predicting the ending embedding. There are different vector operations that can be used for this purpose. Addition operation is one option, however, addition is commutative whereas the relative temporal ordering of the sentences in a story is not. Taking into account the temporal progression of a story, we propose to use the distance vector between adjacent sentences: for a given context of sentences a, b, c, d, we need to predict the distance vector $e - d$ which then predicts the ending vector e. This can be achieved using a basic multivariable curve fitting among the distance vectors of adjacent sentences, e.g., using linear least squares error. Of course the validity of this technique, or any other ones trying to compose the sentence vectors into one vector, requires large scale testing and a comprehensive analysis. As with the other vector space evaluations such as word analogy, further details about this evaluation setup should be finalized after future experiments.

3.3 Baselines

We present preliminary results on evaluating basic embedding models on Story Cloze Test. Here we use the test set split of the available Story Cloze Test dataset, comprising of 1,872 instances. We experiment with the following models:

1. Word2Vec: Encodes a given sentence or paragraph with its average per-word word2vec (Mikolov et al., 2013) embedding.

6. Skip-thoughts Model: A Sentence2Vec embedding (Kiros et al., 2015) which models the semantic space of novels. This model is trained on the 'BookCorpus' (Zhu et al., 2015) (containing 16 different genres) of over 11,000 books. We retrieve the skip-thoughts embedding for the two alternatives and the four sentences, representing the context as the average embedding of the four sentences.

9. Deep Structured Semantic Model (DSSM): This model (Huang et al., 2013) learns to project two different inputs into the same vector space, consisting of two separate embedding modules. It is trained on ROCStories corpus, consisting of 49,255 stories. We retrieve the DSSM embedding for the two alternatives and the context of four-sentences.

For this evaluation we use the joint paragraph and sentence level evaluator module (Section 3.1). Table 2 shows the results, where 'constant' model simply chooses the first alternative constantly. As the results show, there is a wide-gap between human performance and the best performing baseline, making this test a challenging new framework for the community.

	Constant	Word2Vec	Skip-thoughts	DSSM	Human
Test Set	0.513	0.539	0.552	0.585	1.0

Table 2: The preliminary results on Story Cloze Test.

4 Major Characteristics

Our proposed method for representation learning captures the linguistic and semantic property of scripts, which has not been captured by any of the other many existing intrinsic benchmarks. Our method goes beyond capturing human ratings of the similarity of two words or sentences, and towards a more interesting linguistic phenomena of capturing 'what is next', which can potentially affect many other downstream applications.

Our evaluation method is very simple to implement and is based on a high quality resource for accurate evaluation. The human agreement on choosing the right ending of the Story Cloze Test is 100%, making the evaluation schema reliable for making further meaningful progress in the field. Story Cloze evaluation together with the dataset are accurately reproducible by the community. Furthermore, hundreds of thousands of Story Cloze instances can be crowdsourced to non-expert workers in the crowd, making the evaluation scalable.

Although the embeddings models will be trained for the specific application of predicting the ending to a given short story, their impact is

not isolated to narrative understanding since they capture the generic characteristics of a sequence of logically related sentences. Hence, we can hypothesize that the context vector representations which perform well on our method can be used as features in other language understanding and commonsense reasoning tasks, e.g., reading comprehension tests (Hermann et al., 2015; Weston et al., 2015; Richardson et al., 2013) which often require a system to infer additional events given a premise paragraph. Of course, demonstrating that this knowledge is indeed transferable well among different language tasks will be the next step. However, given that the Story Cloze Test is designed as a test of a model's ability to understand and reason with language in a fairly general sense, it does seem plausible that success on Story Cloze Test can translate into success in other downstream language understanding tasks.

5 Conclusion

In this paper we propose a new method for vector representation evaluation which captures a model's capability in predicting what happens next given a context. Our evaluation methodology and the dataset are simple, easily replicable and scalable by crowdsourcing for quickly expanding the resource. Human performs with an accuracy of 100% on this task, which further promises the validity of benchmarking the progress in the field using this evaluation method.

Representation learning community's focus on commonsense reasoning and inferential frameworks can help the research community to make further progress in this crucial area of NLP and AI. We expect the embedding models which somehow leverage commonsense knowledge, perhaps in the form of narrative structures or other knowledge resources, to perform better on our evaluation framework. We believe that a vector representation that achieves a high score according to the Story Cloze Evaluator is demonstrating some level of commonsense reasoning and deeper language understanding.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful comments. This work was supported in part by Grant W911NF-15-1-0542 with the US Defense Advanced Research Projects Agency (DARPA), the Army Research Office (ARO) and the Office of Naval Research (ONR).

References

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1693–1701. Curran Associates, Inc.

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.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. *NIPS*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

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 NAACL HLT*, San Diego, California. Association for Computational Linguistics.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*, pages 193–203. ACL.

Roger C. Schank and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures*. L. Erlbaum, Hillsdale, NJ.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 298–307.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP*.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR*, abs/1502.05698.

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*.