IIT (BHU): System Description for LSDSem'17 Shared Task

Pranav Goel and **Anil Kumar Singh**
Department of Computer Science and Engineering
Indian Institute of Technology (BHU), Varanasi, India
{pranav.goel.cse14, aksingh.cse}@iitbhu.ac.in

Abstract

This paper describes an ensemble system submitted as part of the LSDSem Shared Task 2017 - the Story Cloze Test. The main conclusion from our results is that an approach based on semantic similarity alone may not be enough for this task. We test various approaches and compare them with two ensemble systems. One is based on voting and the other on logistic regression based classifier. Our final system is able to outperform the previous state of the art for the Story Cloze test. Another very interesting observation is the performance of sentiment based approach which works almost as well on its own as our final ensemble system.

1 Introduction

The Story Cloze Test (Mostafazadeh et al., 2016) is a recently introduced framework to evaluate story understanding and script learning. Representation of commonsense knowledge is major theme in Natural Language Processing and is also important for this task. The organizers provide a training corpus called the ROCStories dataset (we will refer to it as the Story Cloze corpus or dataset). It consists of very simple 98161 everyday life stories (combining the spring and winter training sets). All stories consist of five sentences which capture 'causal and temporal common sense relations between daily events'. The validation and test sets contain 1871 samples each, where each sample contains the first four sentences (the context) of the story, and the system has to complete the story by choosing the fifth sentence (the correct ending) out of the two alternatives provided.

Some of the approaches described in (Mostafazadeh et al., 2016) are used as it is in our system, while some approaches not tried before in the context of this task (to the best of our knowledge) also form parts of our final ensemble models. Most approaches tried before and also in our experiments rely directly or indirectly on the idea of using semantic similarity of the context and the ending to make the decision. The results point to the conclusion that semantic similarity (at least on its own) may be inadequate as an approach for the Story Cloze test.

Our final system is an ensemble combining the different approaches we tried. It achieves an accuracy of 60.45 on the test set.

The paper is structured as follows. The next section describes various experiments and approaches we tried. Section 3 describes how the different approaches come together to form the system we submitted. Section 4 looks at the various results and draws inferences to make our point. Section 5 presents a small error analysis. Finally, Section 6 presents the conclusions and discusses possible future work.

2 Approaches

We tried five different approaches, out of which four are directly or indirectly utilizing the idea of semantic similarity between the context and the ending. Some past approaches are mentioned here again to enable readers to view them as semantic similarity based approaches, and to use their performance in our observations and conclusion. We give a brief description of our experiments below. The results (performance measured using accuracy which is simply the correct cases divided by the total number of test cases) of all the separate approaches are presented in Table 1.

1. **Gensim (Average Word2Vec):** Chooses the hypothesis with the closest average word2vec (Mikolov et al., 2013) embedding

Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, pages 81–86,
Valencia, Spain, April 3, 2017. ©2017 Association for Computational Linguistics

to the average word2vec embedding of the context. The concept of semantic similarity is at the very center of this approach. We tried three different variations of this approach:

a) **Training on the Story Cloze training corpus:** This is the same as in (Mostafazadeh et al., 2016) except that we train on the winter training set as well, which makes the corpus size about two times the one used previously. Removing the stop words, keeping a context window of 10 words and vector dimensionality of 300 gave us the results reported in Table 1.
b) **Training on Google news corpus:** Google has released its pre-trained word vectors, trained on a news corpus with a vocabulary of about three million words, which is much larger than the Story Cloze corpus (which contains about 35k unique words). Thus, we decided to explore if the larger set could potentially result in better representation and performance.
c) **Learning the representation of a potential connective word between the context and the ending**: The idea is that a connective with a particular 'sense' (probably temporal or causal in the Story Cloze training set) could perfectly link the context and the ending. We modified all the stories such that a manually introduced symbol (like 'CCC': not in the vocabulary) separates the first four sentences from the fifth sentence, and its representation is learned by training a word2vec model on the data. On the test and validation set, the hypothesis whose representation is the closest to the sum of the vectors of the context and the connective symbol is chosen as the prediction. The intuition comes from the implicit connective sense classification task for the Shallow Discourse Parsing problem (Xue et al., 2015). Context window size 100 and dimensionality 300 were found to be the optimal hyperparameters in our experiments.

Combining the above three – called word2vec (combined) approach – through simple voting produced slightly better results than with any individual variation (as can be seen in Table 1), and thus we used this combined approach in our ensemble model.

2. **Skip-thoughts Model:** The skip-thoughts model's (Kiros et al., 2015) sentence embedding of the context and the alternatives is again compared like the Gensim model, and thus this approach also revolves around semantic similarity.

3. **Gensim Doc2Vec:** Distributed representation of documents and sentences extends the concept of word vectors to larger textual units (Le and Mikolov, 2014). A host of variations were tried (as provided by Python's Gensim functionality (Řehůřek and Sojka, 2010)). The distributed bag of words model (dbow) along with a context window of three words was found to give the best results for this approach (Table 1). This approach is again trying to model semantic similarity via sentence embedding.

4. **Siamese LSTM:** We also implement a deep neural network model for assessing the semantic similarity between a pair of sentences. It uses a Siamese adaptation of the Long Short-Term Memory (LSTM) network (Mueller and Thyagarajan, 2016). The model is implemented as in the paper - using the SICK training set and Google word2vec, with the weights optimized as per the SemEval 2014 task on semantic similarity of sentences (Marelli et al., 2014). This is one of the current state of the art models for capturing semantic similarity.

5. **Sentiment:** In this approach, we choose the hypothesis that matches the average sentiment of the context. We use NLTK VADER Sentiment Analyzer (Hutto and Gilbert, 2014) instead of the Stanford Core NLP tool for sentiment analysis by (Manning et al., 2014) as used in (Mostafazadeh et al., 2016) due to notably better results (Table 1). In our experiments on the validation set, matching sentiment of the full context instead of just the last one/two/three sentence(s) gives the best performance for this approach. This approach does not use semantic similarity.

3 The Ensemble Model

We tried various ways of combining the power of the different approaches, comparing the perfor-

	Gensim word2vec				Skip-thoughts	Gensim doc2vec	Siamese LSTM	Sentiment
	Story cloze	Google word vectors	Using potential connective rep.	*Combined*				
Validation	0.58	0.577	0.571	0.593	0.536	0.547	0.549	**0.608**
Test	0.571	0.568	0.576	**0.584**	0.552	0.546	0.551	0.582
	0.539	-	-	-	0.552	-	-	0.522

Table 1: Results for individual approaches (last row represents results on the test set for corresponding approach in (Mostafazadeh et al., 2016)

	Approaches involving semantic similarity (logistic regression on validation set)	All approaches (includes sentiment)		Baseline
		Weighted majority voting (Final system submission for validation set spring 2016)	Logistic regression on validation set (Final system submission for test set spring 2016)	
Validation	-	**0.626**	-	0.604
Test	0.587	0.601	**0.605**	0.585

Table 2: Results for the best ensemble models

mances of each on the validation set. This creation of an 'ensemble' model was also tried without using the sentiment approach, so as to observe the best possible performance when only our approaches which involve semantic similarity are combined. We report only the best performing combinations (out of all possible combinations of approaches reported above) here:

a) **Voting based ensemble:** We use weighted majority voting, with prediction from sentiment approach counted twice, and predictions from Siamese LSTM, word2vec (combined) and doc2vec counted once each. The idea is to improve the performance of sentiment approach (the best individual performer) by changing its prediction when all the other three approaches predict a different ending. It may be noted that such voting based methods did not lead to improvement (over combined word2vec) when combinations of only semantic similarity based approaches were used.

b) **Applying a supervised machine learning algorithm:** We used the predictions from sentiment, Siamese LSTM and word2vec (combined) approaches on the validation set as features, with the actual validation set labels as targets and train a machine learning classifier on them. Then this classifier predicts the test set labels (with the same set of features created for test set). Logistic regression (C=0.1) gave the best performance in this method (more than decision tree based methods and naive bayes, and also slightly better than SVM for test as well as validation data). This is the system which formed our final submission. Additionally, combining predictions of doc2vec, word2vec, skip-thoughts, and Siamese LSTM in the exact same way gave us the best performance in the case of using only semantic similarity based approaches (see Table 2).

Baseline: We compare our submitted system with the best performing model on the ROCStories dataset for the Story Cloze task (Mostafazadeh et al., 2016) in Table 2.

4 Results and Discussion

We discuss insights and observations gained from the results of our ensemble system and of the individual approaches obtained on the Story Cloze validation and test sets.

1. **Word vectors:** From Table 1, we can see that word vectors on the Story Cloze corpus perform slightly better than the ones pre-trained on Google news corpus, which has a much larger vocabulary (almost 100 times). This shows that the nature or the domain of the training data really matters for this task. So,

further increase in the Story Cloze training data itself may help by giving us better representations. However, comparing with results in (Mostafazadeh et al., 2016), doubling the size of training set results in about 3-4% increase in performance (Table 1). For further increase, trying different approaches might be better.

2. **Improved performance of the sentiment approach:** For the sentiment approach, using NLTK VADER sentiment analyzer tool for getting polarity scores works notably better by outperforming the Stanford Core NLP (Manning et al., 2014) tool used in (Mostafazadeh et al., 2016) by about 6-7% (the last column of Table 1). As discussed in (Hutto and Gilbert, 2014), the VADER tool is about as accurate in most domains and optimal for the social media domain while being quite simple and more efficient. It happens to work surprisingly well in the context of this task though we do not conclude that it is a better approach as compared to (Socher et al., 2013) approach to sentiment analysis as utilized in Stanford Core NLP tool in general.

3. **General performance:** Our best system (ensemble of sentiment and various semantic similarity based approaches) outperforms the previous best system (using DSSM, as given in (Mostafazadeh et al., 2016)) by about 2% (accuracy on both validation and test sets) (refer to Table 2). Most of the individual approaches (Table 1) show performance that hovers around 60% accuracy (or below). Since they are basically all based on semantic similarity (except the sentiment base approach), the results indicate that we may need to approach the Story Cloze test from a very different direction.

4. **Semantic vs. sentiment similarity:** We can see from Table 1 that the simple sentiment based approach basically outperforms all the semantic similarity based approaches. Even combining those approaches seems barely better than just the sentiment approach (Table 2). This could indicate either the lack of effectiveness of semantic similarity or the fact that sentiment based approach is quite effective. Since our sentiment based approach does not rely on training corpus and is unlikely to improve with more data (since no learning is involved), we are inclined towards the former inference: Semantic similarity alone may not be enough for the Story Cloze test.

5. **Negative results of the Siamese LSTM:** Siamese LSTM is a deep neural network trained to capture semantic similarity and gave state of the art results on the data for SemEval 2014 shared task on semantic similarity. However, it does not perform well for this task, supporting our conclusion.

6. **Insignificant boost in performance by ensemble system:** Our final ensemble model (Table 2, last column) offers hardly any improvement over the individual sentiment approach (Table 1, last column). This may indicate that the sentiment and semantic similarity based approaches are not complementary.

5 Error Analysis

Table 3 shows examples where our final ensemble system (the one we submitted for test set) and all the individual approaches (as per table 1) simultaneously chose the wrong ending. We believe that a better understanding of commonsense and a good sense of which alternative is the *logical conclusion* based not only on semantic similarity or sentiment, but the temporal aspect of the chain of events as well as plot consistency is missing. In the first example, the model needs to understand that the first three sentences constitute a ‘prejudice’, and how becoming friends with Sal, who is the target of the prejudice, could lead to the protagonist (Franny) doubting her biased opinion. In the second example, once again, the model would need to understand that the context probably means a nice and happy day for Feliciano, which requires some world knowledge and the sense that spending time like that with a loved one (the grandmother) should lead to happiness. Both the incorrectly chosen endings are inconsistent with the last sentence of the context – Franny being deported – does not make semantic sense when she liked the immigrant, and was not the immigrant herself (we know that the immigrant would get deported and not Franny by our commonsense), while it would not make temporal sense for Feliciano to go picking olives after already collecting them and coming back home to eat with his grandmother.

Context	Incorrect Ending	Correct Ending
Franny did not particularly like all of the immigration happening. She thought immigrants were coming to cause social problems. Franny was upset when an immigrant moved in next door. The immigrant, Sal, was kind and became friends with Franny.	Franny ended up getting deported ˙	Franny learned to examine her prejudices.
Feliciano went olive picking with his grandmother. While they picked, she told him stories of his ancestors. Before he realized it, the sun was going down. They took the olives home and ate them together.	The pair then went out to pick olives.	Feliciano was happy about his nice day.

Table 3: Examples of stories incorrectly predicted by our model as well as all individual approaches

It is interesting to note how the sentiment approach fails in both the examples. NLTK Vader rates ‘getting deported’ as neutral while giving a highly negative rating for ‘prejudice’. The context is only slightly negative, since the positivity in the last sentence (which talks about Sal being ‘nice’ and the act of ‘becoming friends’) offsets the negativity of the previous sentences somewhat. We can see that perhaps the very use of sentiment is not appropriate for example 1. In example 2, the context and the incorrect ending are both neutral, while the correct ending is very positive, hence similarity in sentiment gives an error, but realizing that the context would give rise to a positive ending would have worked.

6 Conclusions and Future Work

We described our submitted system for the Story Cloze test, which combines simple sentiment based approach with a variety of semantic similarity based methods. By highlighting individual and ensemble model results as well as the observations arising from them, we have tried to establish the apparent lack of effectiveness of solely semantic similarity based approaches for this task. This is validated by various experiments and especially the performance of the current state of the art approach for semantic similarity (Siamese LSTM). Also, an effective future approach should probably be more sophisticated than our sentiment based approach, which does not learn from the training data in any way.

We do not claim that semantic similarity or sentiment based approaches are of no help as they may certainly complement the performances of future approaches. However, they do not seem to be enough on their own, though it is certainly possible that some other semantic similarity based models designed for the Story Cloze training set perform better than our approaches.

While word vectors, sentiment based approach and skip-thoughts sentence embeddings had already been discussed as possible approaches before, we also look at two approaches which have not been tried before for this task, namely Siamese LSTM and Gensim Doc2Vec.

For our future work, we plan to build better ensemble methods. Another idea we are keen to try is logical entailment, since the context entails the ending, and a model which can detect this effectively should be able to predict the right ending (our observations of the validation set make it clear that the context would certainly not be entailing a wrong hypothesis).

References

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

Ryan Kiros, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of Advances in neural information processing systems*, pages 3294–3302.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceeding of ICML*, volume 14, pages 1188–1196.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the ACL (System Demonstrations)*, pages 55–60.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceeding of Advances in neural information processing systems*, 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*, pages 839–849.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceeding of AAAI*, pages 2786–2792.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the CoNLL Shared Task*, pages 1–16.