Predicting the Correct Ending of a Story – An Ensemble System

Pranav Goel and Anil Kumar Singh

Indian Institute of Technology (Banaras Hindu University) Varanasi, India

Introduction

Task – The Story Cloze test

Significance – Evaluating story understanding

Training set – ROCStories corpus - very simple 98161 everyday life stories - five sentences which capture 'causal and temporal common sense relations between daily events'.

Validation and Test set – 1871 samples, each containing first 4 sentences (context) of the story, and two alternative endings (the 5th sentence)

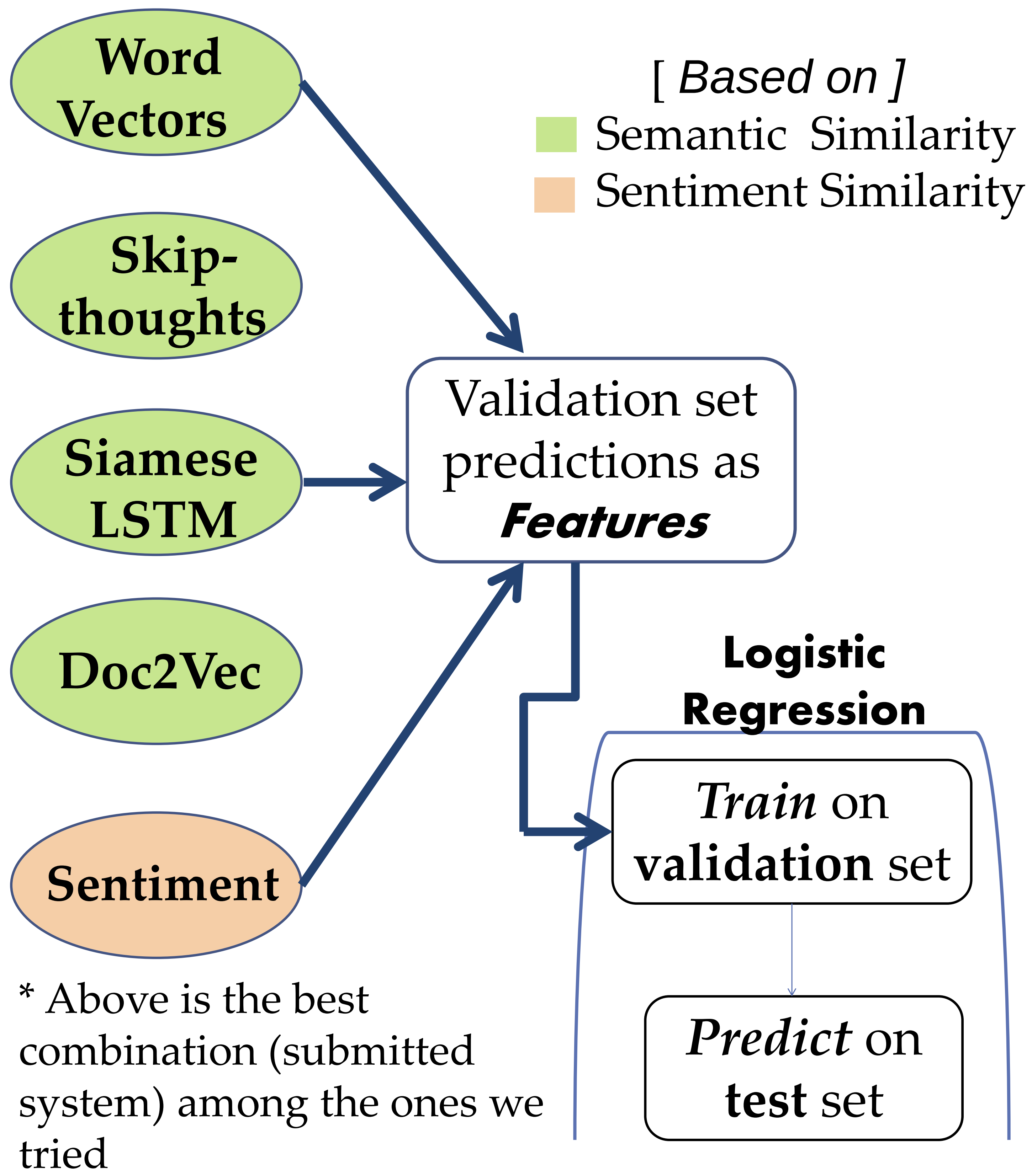
Aim for the system - Choose the correct ending out of the two alternatives.

Sara had lost her cat. She was so sad. She put up signs all over the neighborhood. Then, a wonderful thing happened.

1. Sarah broke her leg.

2. Somebody found her cat

Approaches and the System



* Above is the best combination (submitted system) among the ones we tried

* Voting based ensemble was tried

* Combining only 🟩 approaches was tried

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.

1. The pair then went out to pick olives.

2. Feliciano was happy about his nice day.

Takeaways

- ***Semantic Similarity***

Not enough on its own; center of almost all approaches; mediocre results in our experiments; reference also uses this; ensemble does not really help; sentiment alone better; even LSTM not good enough; new direction!

- ***Sentiment***

Potentially useful; almost beats previous best on its own; more sophisticated exploitation could be useful; only change of tool gives 6-7% jump; does not seem complementary to semantic similarity

- ***Ensemble***

Helps; 2% over previous best; not high improvement over individual; supervised ML based > voting based; only semantic similarity based validates the first point above.

Results *('MZ16' refers to the main reference)*

Approach	Test set Accuracy	MZ16		*What Changed?*
Word vectors	58.4	53.9	Word vectors	(2 X Training data)
Skip-thoughts	55.2	55.2	Skip-thoughts	(No Change)
Siamese LSTM	55.1	--	-	(New)
Doc2Vec	54.6	--	-	(New)
Sentiment	58.2	52.2	Sentiment	(NLTK Vader > Stanford CoreNLP)
Ensemble (only semantic similarity)	58.7	--	-	-
Overall Ensemble (Best)	**60.5**	58.5	DSSM	(Our best versus MZ16's best)

Main Reference

A corpus and evaluation framework for deeper understanding of commonsense stories. – Mostafazadeh et al., 2016 (MZ16)