Story Cloze Task: UW NLP System

Roy Schwartz[1,2], Maarten Sap[1], Ioannis Konstas[1], Li Zilles[1], Yejin Choi[1] and Noah A. Smith[1]

[1]University of Washington, [2]Allen Institute for Artificial Intelligence

Abstract

This poster describes University of Washington NLP's submission for the Linking Models of Lexical, Sentential and Discourse-level Semantics (LSDSem 2017) shared task—the *Story Cloze Task*. Our system is a linear classifier with a variety of features, including both the scores of a neural language model and style features. We report 75.2% accuracy on the task.

Story Cloze Task

Story Prefix	Endings
Joe went to college for art. He graduated with a degree in painting. He couldn't find a job. He then responded to an ad in the paper.	Then he got hired.
	Joe hated pizza.

Approach 1: Language Modeling⁺

$$e^* = \underset{e \in \{e_1, e_2\}}{\operatorname{argmax}} \frac{p_{lm}(e|\text{prefix})}{p_{lm}(e)}$$

Approach 2: Style

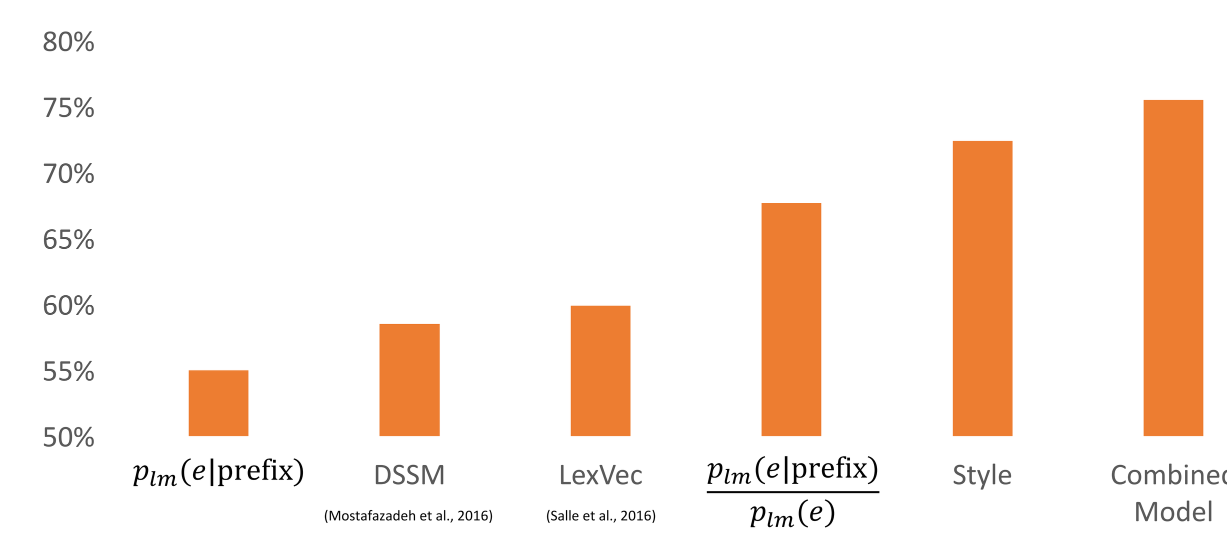
- Intuition: authors use different **style** when asked to write *right* vs. *wrong* story ending
- We train a style-based classifier to make this distinction
- Features are computed using **story endings only**
  - Without considering the story prefix

The brownies are so *delicious* Laverne eats two of them.	**Lina** now knew that candy canes were *boring*.
His boss *commends* him for a job *well done*.	I was very *ashamed* of my performance.
Eventually I *healed*.	I am *dishonest*.
We had *a great time*!	**Ron** started collecting bottle caps.

Combined Model

- A logistic regression classifier
- LM features: $p_{lm}(e|\text{prefix})$, $p_{lm}(e)$, $\frac{p_{lm}(e|\text{prefix})}{p_{lm}(e)}$
  - An LSTM RNNLM trained on the ROC story corpus
- Style features: sentence length, character 4-grams, word 1-5-grams
- Model is trained and tuned on the story cloze development set

Results



Discussion: Language Modeling⁺

$$log\left(\frac{p(e|\text{prefix})}{p(e)}\right) = log\left(\frac{p(e,\text{prefix})}{p(e)p(\text{prefix})}\right) = \boldsymbol{PMI}(e,\text{prefix})$$

Analysis

Freq. ≥ 5%

Right	Freq.	Weight	*Wrong*	Freq.	Weight
'ed .'	6.5%	0.17	START NNP	54.8%	0.21
'and '	13.6%	0.15	NN .	47.5%	0.17
JJ	45.8%	0.14	NN NN .	5.1%	0.15
to VB	20.1%	0.13	VBG	10.1%	0.11
'd th'	10.9%	0.13	START NNP VBD	41.9%	0.11
'lly '	5.0%	0.11	'ecid'	6.5%	0.11
'er .'	5.9%	0.08	NNS .	9.6%	0.10
'for '	6.0%	0.07	'ided'	6.2%	0.10
'ally'	3.3%	0.21	'hate'	1.9%	0.31
VBD the NN .	2.3%	0.21	' hat'	2.0%	0.31
START RB	3.1%	0.21	'ated'	3.0%	0.19
'ved '	4.1%	0.19	'turn'	1.6%	0.17
' tim'	2.6%	0.18	'hrew'	1.2%	0.16

Discussion: Style

- different writing tasks $\xrightarrow{\text{(mental state?)}}$ different writing style
- Common sense induction is hard!
  - Our style-features constitute a strong baseline for the task
  - Our RNNLM is learning something beyond shallow features
- **Schwartz** et al., 2017, *The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task*

Conclusions

- For this task, language models are useful only in the PMI setting
- A style-aware model achieves 72.4% accuracy on the task, without considering the story prefix
- A joint model yield best performing results on the task: 75.2%

roysch@cs.washington.edu http://homes.cs.washington.edu/~roysch/