HorrifAI: Using AI to Generate Two-Sentence Horror

Jennie Chen, Adam Keppler

{jchen437, akeppler} @ stanford.edu

Objective

- Two-sentence horror stories are a form of flash fiction in which a self-contained horror story is told in only two sentences
- **Goal:** Given the first sentence, generate the second sentence of a two-sentence horror story
- **Challenges:** distinctive style, need for a full narrative arc in a limited scope
- **Evaluation targets:** language fluency and story cohesiveness

Dataset

- Stories scraped from subreddit /r/TwoSentenceHorror
- **Preprocessing:**
  - Replaced Unicode characters with ASCII equivalents
  - Filtered out non-story posts and incorrect-length posts
  - Split stories into first and second sentences
- **Generation dataset:** 22,000 examples total
- **Classification dataset** for evaluating story cohesiveness: 44,000 examples
  - Shuffled second sentences between examples
  - 22,000 original ex. = "cohesive", 22,000 shuffled ex. = "not cohesive"
- 85% training, 7.5% validation, 7.5% test for both datasets

Baseline

- Non-neural retrieval baseline using **Jaccard similarity**
  - Ratio of *number of unique words common to both texts* to *number of unique words across both texts*
  - Useful indicator of contextual similarity

$$Jaccard = \frac{|A \cap B|}{|A \cup B|}$$

- Generation: For each example, find and output target sentence from training dataset with highest Jaccard similarity to input sentence

Model Architectures

Transformer

- Based on Vaswani's transformer model [1]
- 6-layer encoders/decoders, 512 hidden units each
- 8 attention heads, scaled dot-product self-attention
- Feed-forward network layer w/ 2048 units, dropout 0.1
- 512-d embeddings + sinusoidal position encoding
- trained with Adam optimizer and NLL loss

LSTM

- 2-layer unidirectional LSTM encoder and decoder with multiplicative attention
- encoder & decoder have 500 hidden units
- 500-d word embeddings, dropout 0.3
- trained with SGD and NLL loss

BLSTM

- Same as LSTM, but uses a bidirectional encoder

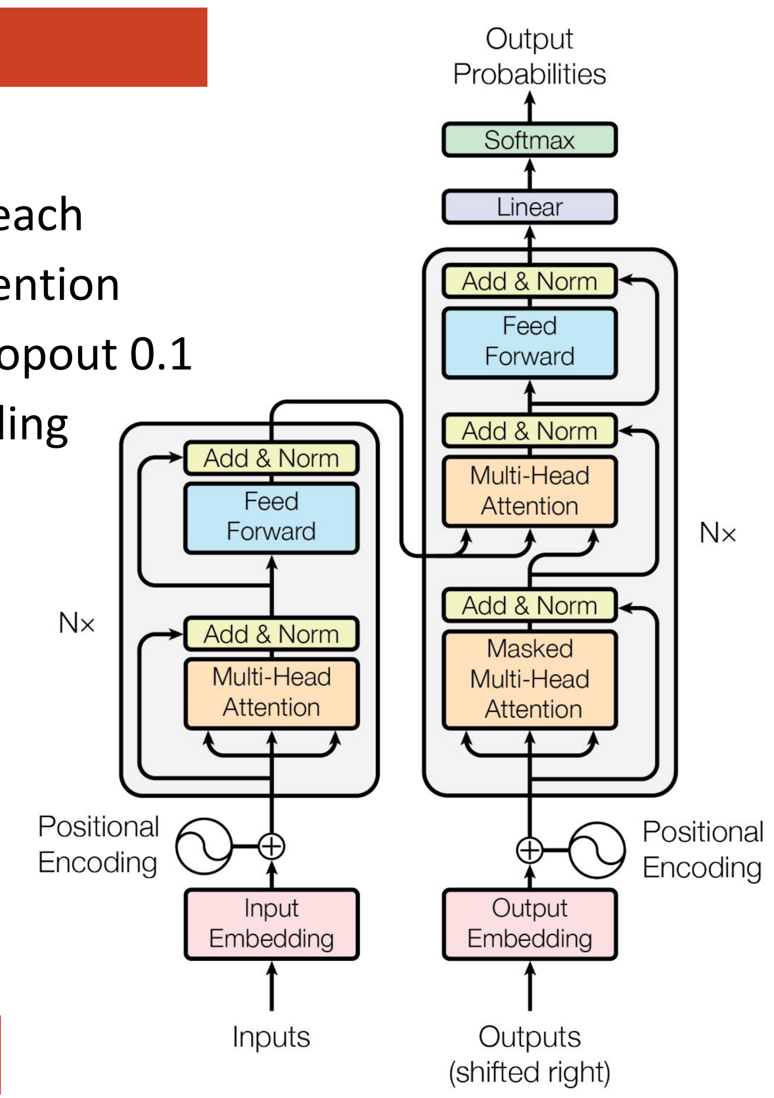

***Figure 1**: Transformer Model [1]*

Evaluation

- All models deliberately overfit to act as retrieval models
- Finding a stopping point for overfitting:
  - **N-gram similarity** - % n-grams in gen. output that exist in train
  - **Exact match** - # gen. outputs that exactly match a sentence in train
- Facebook's InferSent [2]
  - Used for automatic cohesion eval.
  - **BiLSTM sentence encoder with max-pooling + classifier**
  - Encoder trained on GloVe Wiki 2014 300-dim word embeddings
  - Pre-trained on SNLI, then trained on classification dataset using SGD
  - Achieved 78.4/70.1 % train/test acc.
- **Human evaluation** (25-question survey)
  - 5 samples/model, randomly selected
  - Questions on readability, cohesion, "horror", AI detection (rated 1-10)

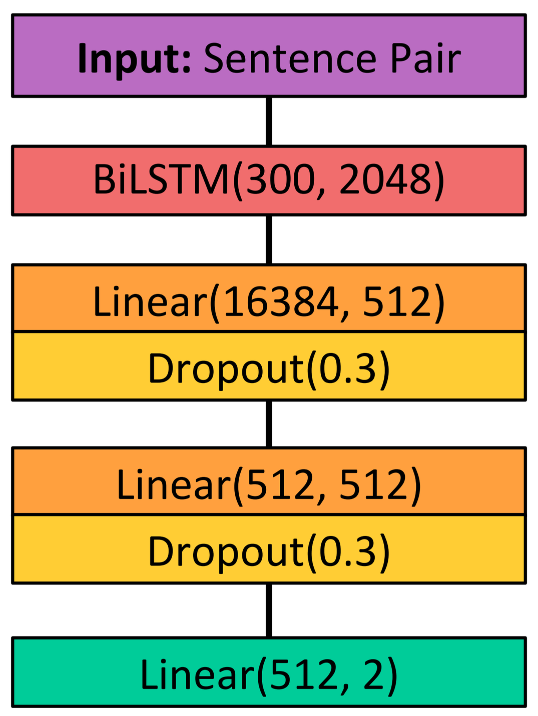

***Figure 2:** Modified InferSent Model*

Decoding Algorithms

- **Random top-k sampling** for k = 3, 5, 8, 10
  - Temperature = 0.5, 0.75, 1.0
  - Higher temperature led to more diversity, but also less readable output
- **Beam search** with beam size = 3, 5, 8, 10
  - Higher beam size yielded better output that was more readable and internally consistent
  - N-gram blocking was not useful

Results

Automated Metrics	2-Gram Sim. %	4-Gram Sim. %	6-Gram Sim. %	# Exact Match	Avg. Jaccard	% Cohesive
Human	*76.92*	*15.55*	*3.31*	*30*	0.068	78.76
Jaccard	*100*	*100*	*100*	*1864*	0.271	85.78
LSTM	94.12	42.78	16.13	166	0.052	61.16
BLSTM	95.03	46.46	17.46	236	0.055	61.00
Transformer	99.68	92.94	89.82	1331	0.046	47.16

***Table 1:** Italicized entries included for completeness, but not very meaningful*

Survey Results	Avg. Readability	Avg. Cohesion	Avg. Quality	% AI vs. Human Acc.
	8.54	8.59	7.69	84.44
	7.14	5.73	5.71	61.48
	6.74	5.21	5.28	68.15
	6.63	4.43	4.63	81.48
	6.93	5.11	5.06	68.88

***Table 2:** Aggregated results from 27 respondents*

Sample Stories

- **Human sample:** *My neighbor's dog dug up some human bones. Guess I gotta start burying them deeper.*
- **Transformer sample:** *I had a staring contest with my reflection. I won.*
- **LSTM sample:** *I really hate when I burn dinner. Thankfully the neighbors have a new dog.*
- **Most mistaken as human-generated (actually Jaccard):** *What can I say, when my life has been a series of farewells, to my family, friends, and all I hold dear? This is all I can remember of the last moments of my life and I'm damned to re-live them over and over and over.*

Can you tell the difference between AI-generated and human-generated stories ? Take our survey: *bit.ly/horrifAI*

Discussion

- Transformer was the best retrieval model and produced readable but not very cohesive output
- Jaccard sim. produced surprisingly cohesive stories
- Using NMT-style seq2seq approach generally effective
  - Output difficult to distinguish from human stories
- Story cohesion is difficult to measure

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2017. arXiv:1706.03762.

[2] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes. Supervised learning of universal sentence representations from natural language inference data. arXiv preprint arXiv:1705.02364, 2017.