Exasperatingly Hard Creative Story Generation

Seraphina Tarrant, Sayan Paul, Nanyun Peng
Information Sciences Institute,
University of Southern California

Challenge

The integral aspects of a high quality story compete with one another.

Aspects: Rel: Relevance Rep: Repetition Cr: Creativity Co: Coherence

Excerpts from baseline architectures:

A. FAIRSEQ (Fan et al, 2018): previous best on this data

✓ **Rel**
X **Rep**
X **Cr**
X **Co**

Prompt: *A Man dies and expects to go either Heaven or Hell , only to be told by an Angel that he already was in Hell.*

I was in Hell. I was in Hell. I was here. I had a job. A job I loved, but it wasn't the work I wanted. The job was a job, a job I worked for. But when I was in Hell, I couldn't stand it anymore.

B. Title-to-Story: our story architecture (diagram, right) without the storyline model

X **Rel**
X **Rep**
✓ **Cr**
✓ **Co**

Prompt: *95 year old man goes to the funeral of the man his deceased wife left him for decades ago.*

His eyes fluttered open, as if to draw a breath. His wrinkled cheeks met sensuously, softly caressed his features as if to reassure him that he was right. His eyelids fluttered open to the ambience, showing soft reds and purples that mingled softly with the dancing light of myriad bulbs.

We try to improve *all* aspects by adapting the Plan-and-Write system; a storyline planning to story generation pipeline (Yao et al 2019). We incorporate discriminators trained on each aspect, as well as global attention on the storyline. We then compare this new system to the above two baselines and a vanilla Plan-and-Write system.

System

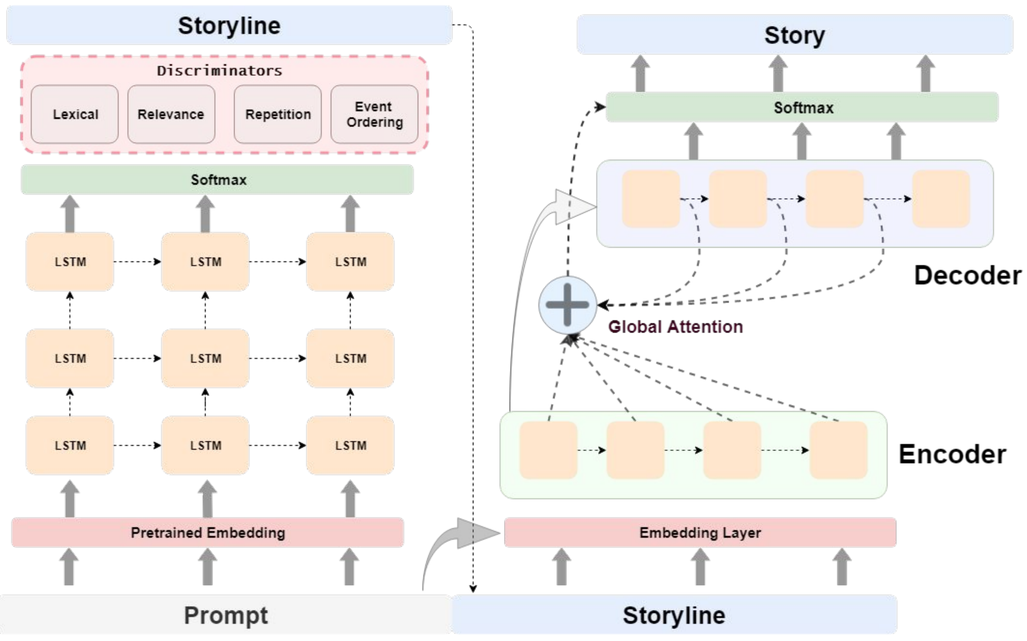


Figure 1: The LSTM language model for storyline generation (left) and encoder-decoder model for story generation (right). Discriminators prefer individual aspects and help extract richer storylines from prompts.

Dataset

WritingPrompts: Prompts and stories from Reddit forum. 272,600 training, 15,138 test and 15,620 validation pairs

Storyline Vocab	96,279
Story Vocab	115,599
Frequency Threshold	10

Metrics

Repetition (objective)

Inter-story (r_e^i) and Intra-story (r_a^i)

$$r_e^i = 1 - \frac{T(\sum_{j=1}^{N} s^{ji})}{T_{all}(\sum_{j=1}^{N} s^{ji})}$$

$$r_a^i = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{\sum_{k=1}^{i-1} T(s^i \cap s^k)}{(i-1) * T(s^i)}\right]^j$$

T = *unique trigrams*
T_{all} = *all trigrams*
s^{ji} = *i^{th} sentence of j^{th} story*
$s^i \cap s^k$ = *trigram intersection between s^i, s^k in one story*

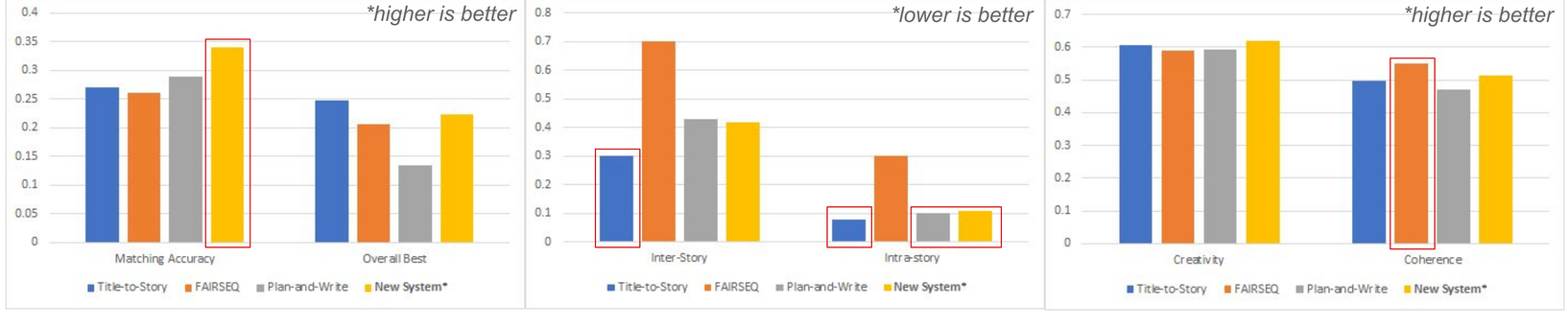
Matching accuracy (objective)

Human judges are asked to match a story with the true prompt out of 3 (two random). *None* is an option, so there is no base % due to chance. Metrics given are $\frac{correct\ pairs}{total\ pairs}$

Creativity, Coherence, Overall (subjective)

Creativity and Coherence are on a 1-5 scale, Overall is % of prompts for which a given system's story was selected as best.

Results



Aggregate results for 105 stories with 3-6 judges per sample. Boxed scores are better with statistical significance < 0.05. As can be seen, the New System always improves upon the Plan-and-Write baseline, but no system is the clear winner across all metrics.