Interactive Open-Domain Story Generation

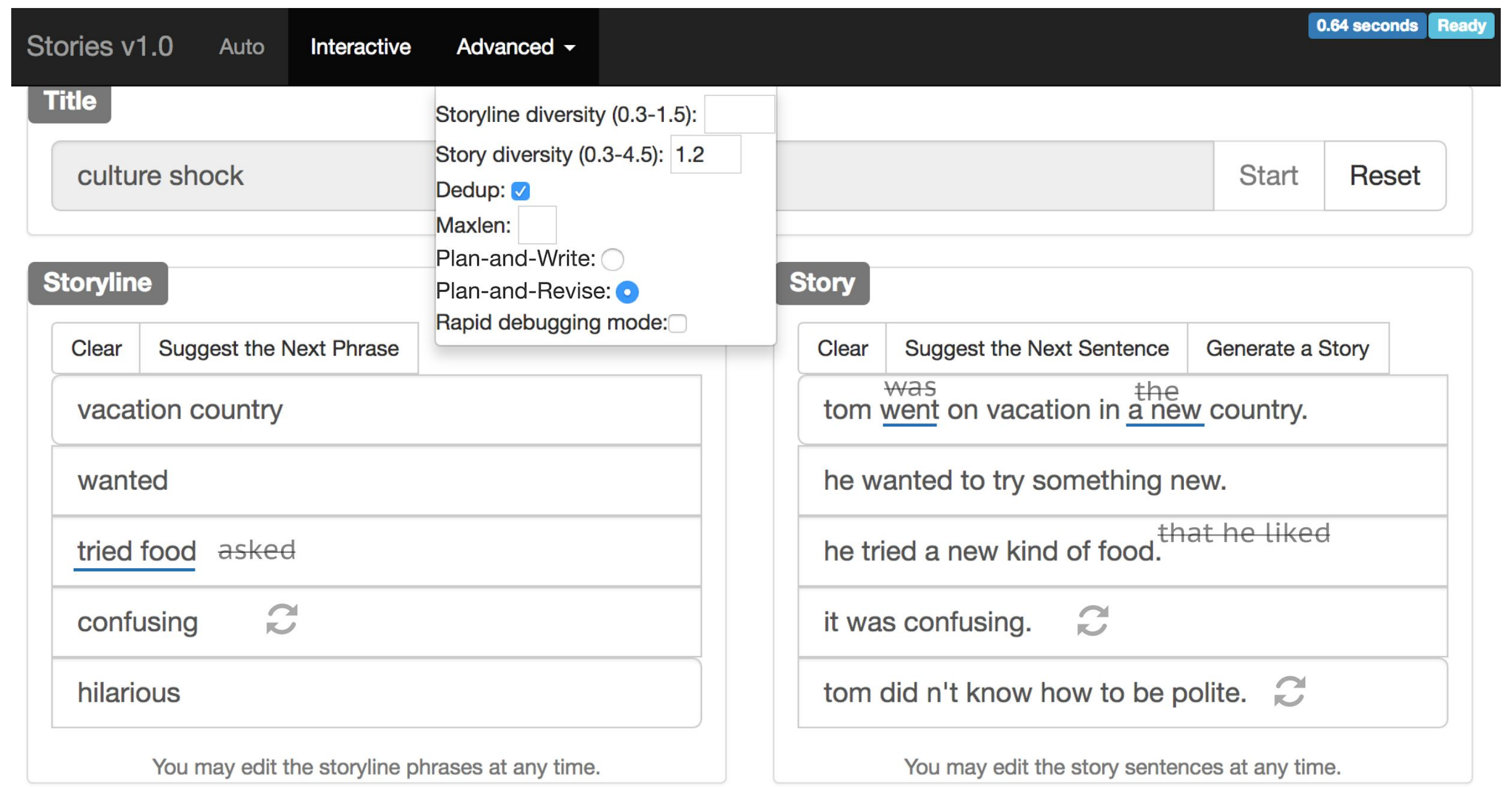
Seraphina Tarrant, Haining Feng, Nanyun Peng
Information Sciences Institute,
University of Southern California

Motivation

- *Can human-machine collaboration improve open-domain neural story generation?*
- *Can it improve specific story aspects, as well as overall quality?*

Previous approaches to human-machine collaboration offer limited interaction. We design a system that enables human interaction at multiple stages of the process: story-planning, story-writing, diversity controls*, and model-selection.

Sample Interaction



Stories v1.0 | Auto | Interactive | Advanced ▾ | 0.64 seconds | Ready

Title: culture shock | Start | Reset

Storyline diversity (0.3-1.5):
Story diversity (0.3-4.5): 1.2
Dedup: ☑
Maxlen:
Plan-and-Write: ○
Plan-and-Revise: ●
Rapid debugging mode: □

Storyline | Clear | Suggest the Next Phrase
- vacation country
- wanted
- tried food ~~asked~~
- confusing
- hilarious

You may edit the storyline phrases at any time.

Story | Clear | Suggest the Next Sentence | Generate a Story
- tom went on vacation in a new country. (~~was~~ above "went", ~~the~~ above "a new")
- he wanted to try something new.
- he tried a new kind of food. ~~that he liked~~
- it was confusing.
- tom did n't know how to be polite.

You may edit the story sentences at any time.

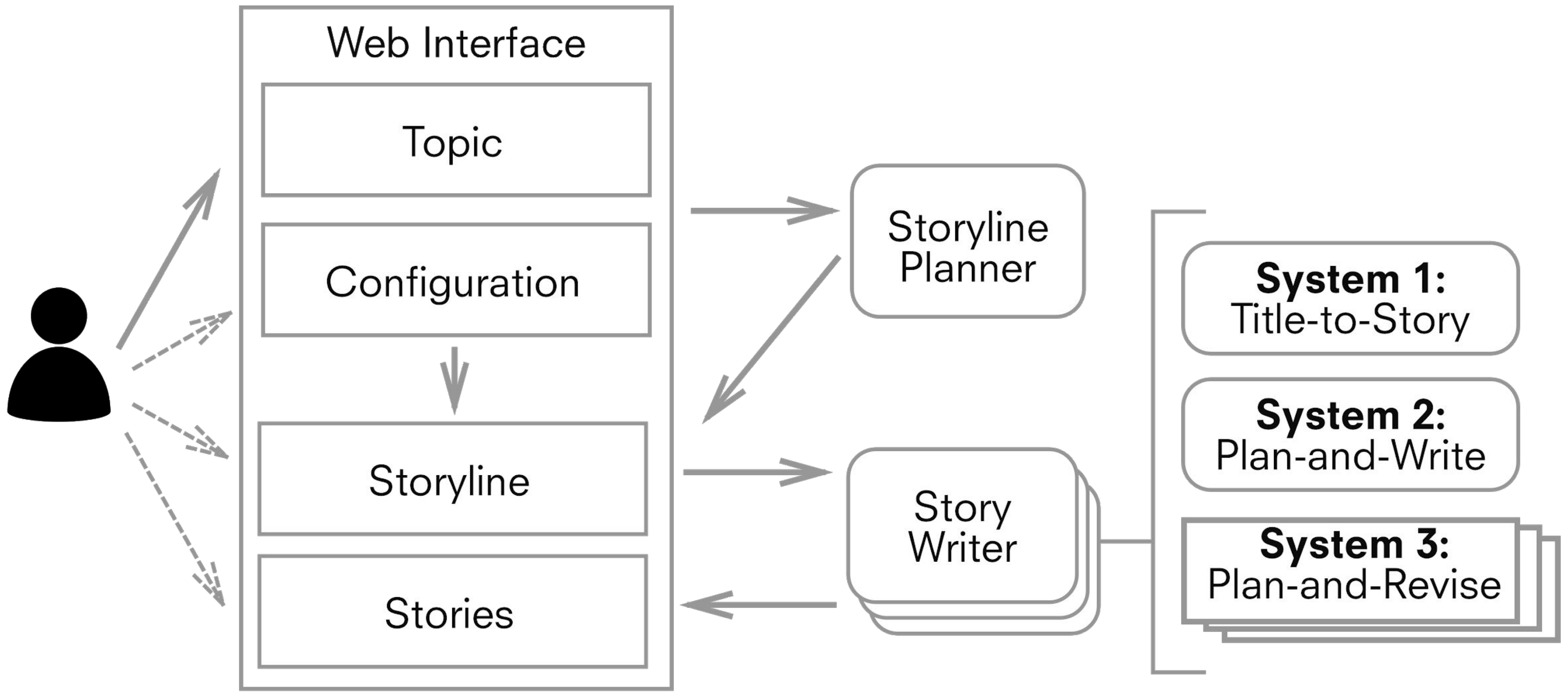
Figure 1: full-interaction capabilities, annotated with user actions from an example study. Interaction is iterative: a user can edit or regenerate any element at any time.

We conduct user studies for multiple interaction scenarios. We constrain experiments to 10 minutes, and explore *full-interaction, story-only, storyline-only,* and *diversity-only* variations.

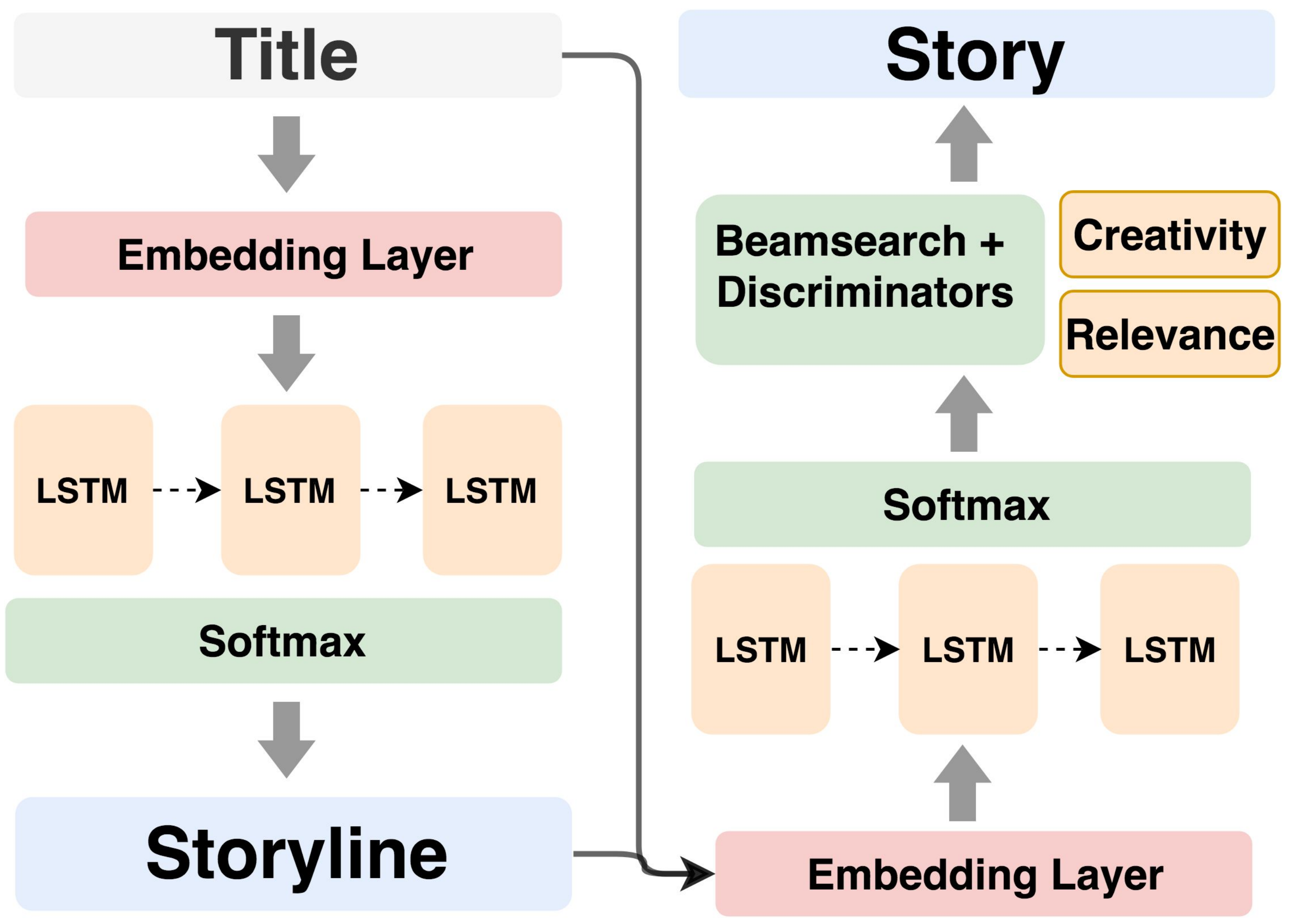
*diversity controls are softmax temperatures, which control the unusualness of system generations

Code and data available at:
https://github.com/seraphinatarrant/plan-write-revise

Demo Diagram



System



Data

ROC Stories: 98,162 commonsense stories data split into 8:1:1 for training, dev and test sets. Storylines are extracted via RAKE (a keyword extraction algorithm) as in Yao et al (2019).

We adapt the Plan-and-Write system; a storyline planning to story generation pipeline (Yao et al 2019) to enable interaction at the story-planning stage. We include their Title-to-Story baseline (no planning stage) and create a new Plan-and-Revise system, which incorporates two discriminators for *Relevance* and *Creativity*, as in Holtzman et al. (2018).

Metrics

Self-reported (subjective)
Subjects self-report on their engagement, satisfaction with their story, and perception of story quality.

Independent Ranking
Independent human judges are asked to rank all stories from 1-5 under eight experiment conditions for Overall Quality, Relevance, Creativity, and Causal-Temporal Coherence.

Results

Experiment	Overall	Creative	Relevant	C-T
Machine	2.34	2.68	2.46	2.54
Diversity only	2.50	2.96	2.75	2.81
Storyline only	3.21	3.27	3.88	3.65
Story only	3.70*	**4.04**	3.96*	**4.24**
All	3.54	3.62	3.93*	3.83
All + Creative	**3.73**	3.96*	3.98*	3.93*
All + Relevant	3.53*	3.52	**4.05**	3.91*
All + C-T	3.62*	3.88*	4.00*	3.98*

Table 1: Results for all experiments. Best scores per metric are bolded, scores not significantly different ($\alpha = 0.1$, per Wilcoxon Signed-Rank Test) are starred. C-T stands for Causal-Temporal Coherence, the **+** experiments are the extensions where the user focuses on improving a particular quality.

- humans tasked with improving a specific story aspect are successful at doing so
- interaction at both *planning* and *writing* stages improves story quality 10-50% over the less interactive baselines.
- additional interaction increases user self-reported satisfaction.