Creative Generation as Inverse Summarization

Nanyun (Violet) Peng
Nov 4th, EMNLP Workshop on New Frontiers in Summarization
Automatic Text Summarization

Source Text: Peter and Elizabeth took a taxi to attend the night party in the city.

While in the party, Elizabeth collapsed and was rushed to the hospital.

Summary: Elizabeth was hospitalized after attending a party with Peter.

Figure credit: Prakhar Ganesh
Creative Story Generation

Can you tell a story about a race?

Sam was a star athlete. He ran track at college. There was a big race coming up. Everyone was sure he would win. Sam got first place.

Can you tell a sad story?
Creative Pun Generation

- **Input:** hare, hair
- **Output:** The greyhound stopped to get a hare cut.
Creative Text Generation is Challenging

- **Title**: bicycle path accident
- **Generated Story**: sam bought a new bicycle. his bicycle was in an accident. his bicycle was in an accident. his bicycle was in an accident. his bicycle was totaled.
How About GPT-2?

- **bicycle path accident**, but it's still a pretty bad event. But, you know, they're also doing other events. ... I'm pretty excited, actually. I'm really excited. And I hope, you know, there's some other people out there feeling it. ...
The Missing Ingredient for Creative Text Generation

• The contents are underspecified
  – How to generate *creative* and *coherent* contents?

• Analyzing/summarizing stories, hierarchical generation.
  – Generate a plot/summary first, then generate the story

• What’s in a story?
  – Characters, genre, key events, morals, conflicts, ending spirit...
From Analyzing to Creative Generation

• Analyze large corpus of creative contents (stories/poems/puns) to
  – Understand:
    • help humans to come up with concrete theories of story, humor in pun, etc.
  – Generate:
    • coarse-to-fine hierarchical generation
    • novel contents that are coherent, creative, and interesting ....
Outline

• Plan-and-Write Hierarchical Story Generation
  – Plan-and-Write (Yao et al. 2019)
  – Controllable generation (Peng et al. 2018)
  – Collaborative story generation (Goldfarb-Tarrant et al. 2019)

• Analyze Event Temporal Relations in Stories
  – Event Temporal Relation Extraction (Han et al. 2019a)
  – Joint Event and Temporal Relation Extraction (Han et al. 2019b)
Analyzing to Generate Stories

- Analyzing stories to generate stories with minimal or no supervision.

**Storyline:** tom → wanted → decided → practiced → won

**Ending Valence:** happy ending, sad ending
Plan-and-Write Hierarchical Generation

• Can computer generate storylines automatically (given titles)?
  – Mimic human writers’ common practice of writing a sketch/plot: have a big picture.
  – Equip our system with the ability to model “what happens next”.
  – Computer and human can interactively modify the storylines, more fun interactions.

Title: christmas shopping
Story: frankie had christmas shopping to do.
she went to the store.
inside, she walked around looking for gifts.
soon her cart was full.
she paid and took her things home.

Storyline (unsupervised extraction): frankie -> store -> gifts -> cart -> paid

Title: farm
Story: bogart lived on a farm.
he loved bacon.
he decided to buy a pig.
shortly after, he grew fond of the pig.
bogart stopped eating bacon.

Storyline (unsupervised extraction): farm -> bacon -> decided -> pig -> bogart
• Dataset: ROCStories – 90k turker generated five-line stories.
• Extraction tool: A modification to RAKE (rapid automatic keyword extraction), Rose et. al. 2010.
• Storyline composition: extracting one word/phrase per sentence, and order them according to the narrative order.
The **planning** component generates storylines from titles. The **writing** component generates stories from storylines and titles.

Tina made spaghetti for her boy friend.
Dynamic and Static Schemas

Dynamic Schema

We define context as: $\text{ctx} = [t; s_{1:i-1}]$

At the plan step, we model: $P(l_i|\text{ctx}, l_{1:i-1})$

At the write step, we model: $P(s_i|\text{ctx}, l_{1:i})$

The probabilities are computed by some specifically designed fusion-RNN cells.

Static Schema

At the plan step, we model: $P(l_i|t, l_{i-1})$

At the write step, we model: $P(s_i|\text{ctx}, l_{1:5})$

The probabilities are computed by standard language models and sequence to sequence with attention models.
Some Observations

• Plan-and-Write strategies generate more interesting, less repetitive stories.
• Plan-and-Write strategies generate more on-topic stories.
• Static strategy works better than dynamic strategy.
Title: gymnastics

Without Storyline Planning

Story (generated):
i wanted to learn how to draw.
so, i decided to go to the gym.
i went to the local gym.
i got a lot of good grades.
i was very happy.

With Storyline Planning

Storyline (generated): wanted -
-> decided -
class -
-> practiced -
-> well

Story (generated):
i wanted to be a gymnast.
i decided to learn how to do gymnastics.
i decided to take a class.
i practiced every day.
i was able to do well on the class.
Generation Results (Cont.)

Title: build a house

Without Storyline Planning

Story (generated):
When I was young, I wanted to build a house. I went to the store and bought all the supplies I needed. I bought the supplies and went home.

With Storyline Planning

Storyline (generated): build house -> decided -> built -> finished -> proud

Story (generated):
john wanted to build a house. he decided to build a house. he built a small house. he finished the house. he was proud of himself.
Quantitative Results on Repetition

Inter- and intra-story tri-grams repetition rates, the lower the better. We also conduct the same computation for four and five-grams and observed the same trends. As reference points, the whole story repetition rates on the human-written training data are 34% and 0.3% for the inter- and intra-story measurements respectively.

![Inter-story repetition rates](image1)

![Intra-story repetition rates](image2)
## User Preferences

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Dynamic v.s. Inc-S2S</th>
<th>Static v.s. Cond-LM</th>
<th>Static v.s. Dynamic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fidelity</td>
<td>35.8%</td>
<td>12.9%</td>
<td>38.5%</td>
</tr>
<tr>
<td>Coherence</td>
<td>37.2%</td>
<td>28.6%</td>
<td>39.4%</td>
</tr>
<tr>
<td>Interesting</td>
<td>43.5%</td>
<td>26.7%</td>
<td>39.5%</td>
</tr>
<tr>
<td>Overall</td>
<td>42.9%</td>
<td>27.0%</td>
<td>40.9%</td>
</tr>
</tbody>
</table>

The human study is conducted on Amazon Mechanical Turk (AMT). 233 users were participated in the study.
In Addition To Hierarchical Generation...
Controllable Generation

Creative Story Generation with Ending Valence Control

Label: *HappyEnding or *SadEnding

Body: four sentences

Ending: the last sentence

Peng et al. (NAACL 2018, Storytelling Workshop).
Annotating Story Endings Valence

- **Dataset:** ROCstories
- **Pilot study:** two researchers, 150 stories; *inter-annotator agreement: 81%*
- **Second round:** designed new guideline. New *inter-annotator agreement: 95%,* excluding the cases that one cannot decide.
- **Four rounds of AMT annotations.** Overall *agreement of turkers’ v.s. researchers’ annotation: 78%.*
- **Training a classifier for the story ending spirit.**
  - BiLSTM feature learning + mean pooling + Logistic Regression classifier.
  - *4000 annotated stories*, 5-fold cross-validation, *classification accuracy: 68.87%*
Training a Conditional Language Model

Annotated Story Corpus

Happy Ending Representation

Sad Ending Representation

Conditional RNN Language Model

- Target word
- "is"
- "the"
- "problem"

- Output likelihood
- Hidden state
- Input embedding
- Input word
- "What"
- "is"
- "the"

- $W_{hy}$
- $W_{hh}$
- $W_{xh}$
Interactive Generation Task: Flipping

An existing story

Label: *HappyEnding

Body: four sentences

Ending: the last sentence

A new story

Label: *SadEnding

Body: the same four sentences as the existing story

Ending: a new sentence which is sad
**Generation Samples**

Quantitative results see Peng et al. (NAACL 2018, Storytelling Workshop).

<table>
<thead>
<tr>
<th>Computer generated</th>
<th>Human flip the *EndingSpirit</th>
</tr>
</thead>
<tbody>
<tr>
<td>*HappyEnding:</td>
<td>*HappyEnding -&gt; *SadEnding:</td>
</tr>
<tr>
<td>i was playing soccer with my friends.</td>
<td>i was playing soccer with my friends.</td>
</tr>
<tr>
<td>we were playing in the park.</td>
<td>we were playing in the park.</td>
</tr>
<tr>
<td>i was a bit nervous.</td>
<td>i was a bit nervous.</td>
</tr>
<tr>
<td>i never played.</td>
<td>i never played.</td>
</tr>
<tr>
<td><em>i ended up winning the game.</em></td>
<td><em>i was n't very good at it.</em></td>
</tr>
<tr>
<td>*SadEnding:</td>
<td>*SadEnding -&gt; *HappyEnding:</td>
</tr>
<tr>
<td>tom was on vacation.</td>
<td>tom was on vacation.</td>
</tr>
<tr>
<td>he was going to the beach.</td>
<td>he was going to the beach.</td>
</tr>
<tr>
<td>he had to go to the beach.</td>
<td>he had to go to the beach.</td>
</tr>
<tr>
<td>he was on the beach.</td>
<td>he was on the beach.</td>
</tr>
<tr>
<td><em>he was feeling sick.</em></td>
<td><em>he was so excited to go back.</em></td>
</tr>
</tbody>
</table>
An Application of Controllable Creative Generation...
Human-Computer Collaborative Generation

Goldfarb-Tarrant et al. (NAACL 2019, Demo).
Augmented Architecture: Plan-and-write + Learning to Write

Holtzman et al. ACL 2018

the not so haunted house

Bobby and his friends were fascinated by the dark. They dared each other to get close to a haunted house...

Discriminators that are helpful on ROCstory dataset: Relevance and Lexical style. Repetition discriminator is not helpful because the plan-and-write framework has already significantly reduced the repetition.

The entailment discriminator is not helpful probably because of the shift of domains.
**User Study**

- **Settings:**
  - **Machine only:** no human-in-loop.
  - **Diversity only:** user can compare and select models but only diversity (sampling temperature) is modifiable.
  - **Storyline only:** user collaborates on storyline but not story.
  - **Story only:** user collaborates on story but not storyline.
  - **All (+ aspect):** user can modify everything. + aspect means users are tasked with improving a specific aspect of a story.
  - **Turn-taking:** user and machine take turns writing a sentence each (user starts). User can edit the machine-generations, but once they move on to later sentences, previous sentences are read-only.
## User Study Results

<table>
<thead>
<tr>
<th>Experiment</th>
<th>E</th>
<th>Q</th>
<th>S</th>
<th>Use Again</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diversity only</td>
<td>3.77</td>
<td>2.90</td>
<td>3.27</td>
<td>1.40</td>
</tr>
<tr>
<td>Storyline only</td>
<td>4.04</td>
<td>3.36</td>
<td>3.72</td>
<td>1.27</td>
</tr>
<tr>
<td>Story only</td>
<td>4.50</td>
<td>3.17</td>
<td>3.60</td>
<td>1.60</td>
</tr>
<tr>
<td>All</td>
<td>4.41</td>
<td>3.55</td>
<td>3.76</td>
<td>1.55</td>
</tr>
<tr>
<td>All + Creative</td>
<td>4.00</td>
<td>3.27</td>
<td>3.70</td>
<td>1.70</td>
</tr>
<tr>
<td>All + Relevant</td>
<td>4.20</td>
<td>3.47</td>
<td>3.83</td>
<td>1.57</td>
</tr>
<tr>
<td>All + C-T</td>
<td>4.30</td>
<td>3.77</td>
<td>4.30</td>
<td>1.53</td>
</tr>
<tr>
<td>Turn-taking</td>
<td>4.31</td>
<td>3.38</td>
<td>3.66</td>
<td>1.52</td>
</tr>
</tbody>
</table>

User self-reported scores, from 1-5. *E*: *Entertainment* value, *Q*: *Quality* of Story, *S*: *Satisfaction* with Story. C-T stands for Causal-Temporal Coherence. The final column Use Again is based on converting “no” to 0, “conditional” to 1, and “yes” to 2.
Collaboration improve story quality

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Overall</th>
<th>Creative</th>
<th>Relevant</th>
<th>C-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine</td>
<td>2.34</td>
<td>2.68</td>
<td>2.46</td>
<td>2.54</td>
</tr>
<tr>
<td>Diversity only</td>
<td>2.50</td>
<td>2.96</td>
<td>2.75</td>
<td>2.81</td>
</tr>
<tr>
<td>Storyline only</td>
<td>3.21</td>
<td>3.27</td>
<td>3.88</td>
<td>3.65</td>
</tr>
<tr>
<td>Story only</td>
<td>3.70*</td>
<td>4.04*</td>
<td>3.96*</td>
<td>4.24*</td>
</tr>
<tr>
<td>All</td>
<td>3.54</td>
<td>3.62</td>
<td>3.93*</td>
<td>3.83</td>
</tr>
<tr>
<td>All + Creative</td>
<td>3.73*</td>
<td>3.96*</td>
<td>3.98*</td>
<td>3.93*</td>
</tr>
<tr>
<td>All + Relevant</td>
<td>3.53*</td>
<td>3.52</td>
<td>4.05</td>
<td>3.91*</td>
</tr>
<tr>
<td>All + C-T</td>
<td>3.62*</td>
<td>3.88*</td>
<td>4.00*</td>
<td>3.98*</td>
</tr>
<tr>
<td>Turn-taking</td>
<td>3.55*</td>
<td>3.68</td>
<td>4.27*</td>
<td>3.81</td>
</tr>
</tbody>
</table>

External rating of the generated stories, from 1-5. Best scores per metric are bolded, scores not significantly different (per Wilcoxon Signed-Rank Test) are starred.

Outline

• Plan-and-Write Hierarchical Story Generation
  – Plan-and-Write (Yao et al. 2019)
  – Controllable generation (Peng et al. 2018)
  – Collaborative story generation (Goldfarb-Tarrant et al. 2019)

• Analyze Event Temporal Relations in Stories
  – Event Temporal Relation Extraction (Han et al. 2019a)
  – Joint Event and Temporal Relation Extraction (Han et al. 2019b)
Temporal Relation Extraction Motivation:

Improve Story Generation:

- Better modelling the event temporal relation via a *planning* stage (plan-and-write)
  - **Title**: broken ankle
  - **Storyline**: walking school ➔ stepped ➔ landed ground ➔ broke ➔ hospital
Plot Graph

- Complex event and temporal relations.
- Forming a graph structure.
- Capture the high-level plot of

Figure credit: Mark Riedl
Event Temporal Relation Extraction

• Task A: Detect events in text
  – Kelly and her friends **decided** to have a hot dog contest. The girls **competed** against each other.

• Task B: Given A, predict temporal / causal relations between two **events** in some contexts
  – Temporal relations: before, after, simultaneous, vague
  – **Example:**
    • Relation \(<e_1: \text{decided}, e_2: \text{competed} > \xrightarrow{\text{before}} \)  
  – The relations between multiple events form a graph just like the plot graph.
End-to-End Event and Temporal Relation Extraction

A pipeline model:

A joint model:
Global Models

- A deep structured support vector machine (Deep SSVM)
  - A *shared representation learner* that extracts features for both events and temporal relations.
  - A *structured prediction* model to make *joint* predictions.

Axioms as equations capture constraints

- Event-relation compatibility constraint
- All event-event relations have an inverse, e.g. cause & caused_by
- Transitivity is captured
Global Models Formulation

● We propose a deep structured support vector machine (Deep SSVM) model to capture constraints in the outputs.

\[
\mathcal{L} = \sum_{n=1}^{l} \frac{1}{M^n} \left[ \max (0, \Delta(y^n, \hat{y}^n) + S(\hat{y}^n; x^n) - S(y^n; x^n)) \right] + C ||\Phi||^2
\]

○ Binary variables \( y_{i,jr} \in \{0,1\} \) indicate whether token i and j has relation r.

○ Symmetric and transitivity constraints

\[\forall (i,j), (j,k) \in \mathcal{E}, y_{i,j}^r = y_{j,i}^r, \text{ (symmetry)}\]

\[y_{i,j}^{r_1} + y_{j,k}^{r_2} - \sum_{r_3 \in \text{Trans}(r_1,r_2)} y_{i,k}^{r_3} \leq 1, \text{ (transitivity)}\]
Event Relation Extraction Results


Results on Three Benchmark Datasets

- **MATRES**
  - Current SOTA: 69
  - Our Local: 80.3
  - Our Global: 81.7

- **TCR**
  - Current SOTA: 71.1
  - Our Local: 79.5
  - Our Global: 80.9

- **TBDense**
  - Current SOTA: 57.7
  - Our Local: 62.5
  - Our Global: 63.2

SOTA for TCR: Ning et al. ACL 2018a.
SOTA for MATRES: Ning et al. ACL 2018b.
SOTA for TB-Dense: Meng and Rumshisky. ACL 2018.
Results on Joint Event and Relation Extraction on TB-Dense


Our variations (neural-based):
- Single-task model + pipeline prediction
- Multi-task model + pipeline joint training
- Multi-task model + pipeline prediction
- Multi-task model + SSVM joint training

Han et al. EMNLP 2019.
Results on Joint Event and Relation Extraction on MATRES

Han et al. EMNLP 2019.

<table>
<thead>
<tr>
<th>Event</th>
<th>Baseline</th>
<th>STM+PL</th>
<th>MTM+PL</th>
<th>MTM+PLJ</th>
<th>MTM+SSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>85.2</td>
<td>86.9</td>
<td>86.4</td>
<td>87.2</td>
<td>87.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Relation</th>
<th>Baseline</th>
<th>STM+PL</th>
<th>MTM+PL</th>
<th>MTM+PLJ</th>
<th>MTM+SSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>52.8</td>
<td>57.2</td>
<td>58.7</td>
<td>58.5</td>
<td>59.6</td>
</tr>
</tbody>
</table>


Our variations (neural-based):
- Single-task model + pipeline prediction
- Multi-task model + pipeline joint training
- Multi-task model + pipeline prediction
- Multi-task model + SSVM joint training
Conclusion

• Creative generation can be viewed as inverse summarization.
  – The under specification of contents bring many challenges.
  – Plan-and-write hierarchical story generation improves coherence and interestingness
  – We explored several fun human-computer interaction setup.
• Event causal and temporal relations modeling can help story generation.
  – We proposed a deep SSVN model for it.
• Code and demo can be found at:
  • https://violetpeng.github.io/
Thanks!

Questions?

Special thanks to my students and collaborators:
Rujun Han, I-Hung Hsu, Seraphina Goldfarb-Tarrant, Haining Fei, Lili Yao, Mu Yang, Qiang Ning, Rui Yan, Marjan Ghazvininejad, Jonathan May, Aram Galstyan, Ralph Weischedel, Kevin Knight