Carnegie Mellon University

PINEAPPLE: Personifying INanimate Entities by Acquiring Parallel Personification data for Learning Enhanced generation

COLING 2022

Sedrick Scott Keh, Kevin Lu, Varun Gangal*, Steven Y. Feng*, Harsh Jhamtani, Malihe Alikhani, Eduard Hovy

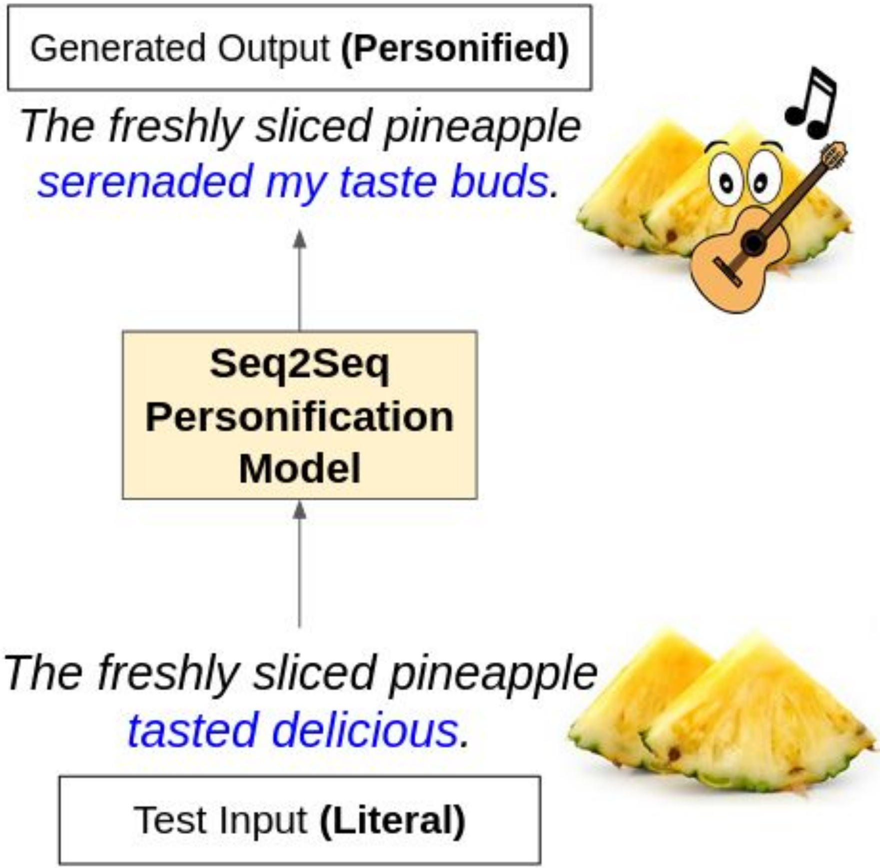
How do we generate coherent, diverse, and interesting personifications?

Important for:

- Dialogue systems
- AI-assisted creative writing

Challenges:

- No explicit structure (unlike similes which use ‘like’ or ‘as’)
- Not as loosely defined as metaphors
- Require model to understand the concept of **<u>animacy</u>**



Task

Given a literal sentence, convert the sentence to a sentence containing a personification.

PersonifCorp Dataset

- 511 diverse personifications
- Gathered from various sources:
  - *CL Prior Art (e.g. Deja Image Captions dataset (Chen et al., 2015))
  - Kaggle/SemEval tasks (e.g. http://www.kaggle.com/datasets/varchitalalwani/figure-of-speech)
- Test set: Human-annotated list of (literal, personification) pairs

In order to train such a model, we will need personification+literal training pairs.

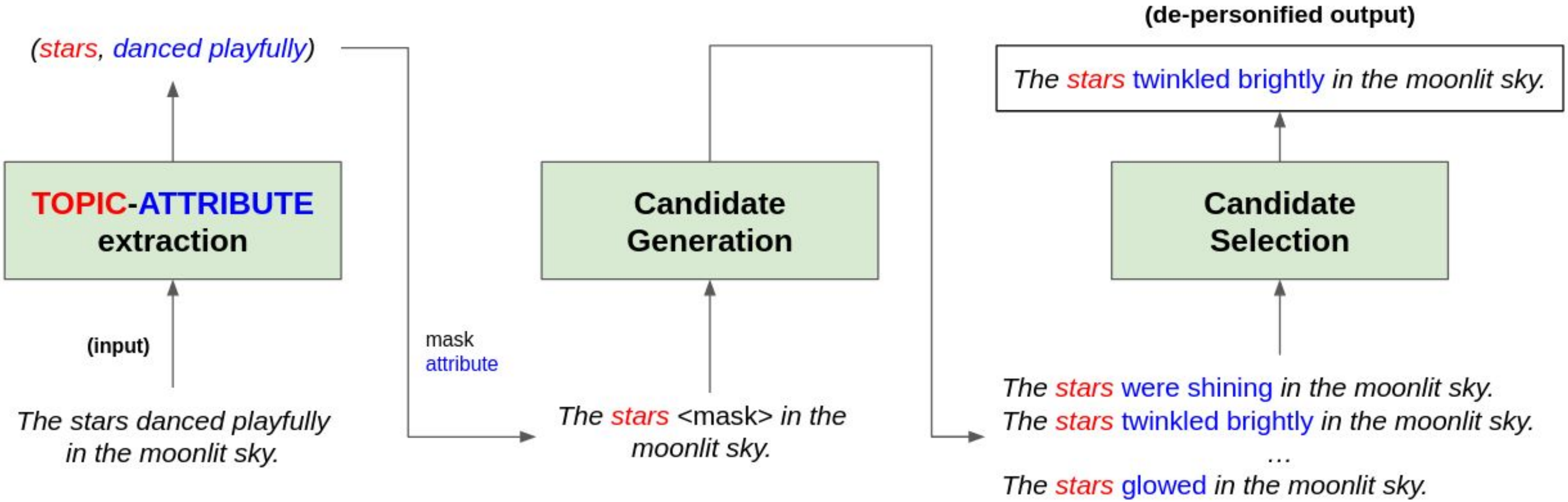
Original Personification	Result After De-Personifying
How far that little candle throws its beams!	How far that little candle can spread its beams!
A book is a fragile creature, it suffers the wear of time, it fears rodents, the elements and clumsy hands.	A book is fragile, it can break from the wear of time, it can be eaten by rodents, the elements and clumsy hands.
The camera loves her since she is so pretty.	The camera is always on her since she is so pretty.
Any trust I had for him walked right out the door.	Any trust I had for him had gone right out the door.
The full moon peeped through partial clouds.	The full moon was visible through partial clouds.
Opportunity was knocking at her door.	Opportunity was knocking at her door.
The killing moon will come too soon.	The killing moon will be here too soon.

Table 1: Example outputs of the **PINEAPPLE** de-personification pipeline. The ATTRIBUTES are highlighted in blue for both the original personifications, as well as the de-personified output sentences. The last two rows contain negative examples where the process does not successfully de-personify the input.

Given our *PersonifCorp* dataset of personifications, how do we "de-personify" a sentence?

Automatic Parallel Corpus Creation

We “de-personify” the personifications using the pipeline below.



Automatic Parallel Corpus Creation

1. TOPIC-ATTRIBUTE Extraction

TOPIC = a noun phrase that acts as a logical subject

ATTRIBUTE = the distinctly animate action or characteristic that is being ascribed to the TOPIC

Dependency parse trees + iterative merging algorithm to determine the TOPICs and ATTRIBUTEs of a given sentence.

ATTRIBUTE Type	Example
Noun	The planet earth is our mother.
Verb	My alarm clock yells at me to get out of bed every morning.
Adjective	Justice is blind and, at times, deaf.

Figure 2: Examples of different types of personification ATTRIBUTES (TOPICS in red and ATTRIBUTES in blue).

Automatic Parallel Corpus Creation

2. Candidate Generation

She did not realize that opportunity was knocking on her door.

(She, did not realize) – animate (we ignore)
(opportunity, knocking on her door) – inanimate

Use COMET's (Bosselut et al., 2019) ConceptNet relations (Speer et al., 2017) as a proxy for animacy.

- IsA(x, "person")

She did not realize that opportunity was <mask>.
Top k=10 candidates:

- "knocking at her door"
- "present"
- "lost"
- …
- "going to arrive"

Use pre-trained BART to generate k=10 candidates for each inanimate TOPIC.

Automatic Parallel Corpus Creation

3. Candidate Selection

Given k=10 replacement candidates, design a ranking system to select the most appropriate candidate:

1. **Animacy** – $\mathcal{A}_{human,\mathrm{ATT}} - \mathcal{A}_{\mathrm{TOPIC},\mathrm{ATT}}$ where $A_{(X,\ ATT)}$ is the COMET CapableOf score between X and the ATTRIBUTE
2. **Fluency** – use BART's generation scores (sum of individual token logits)
3. **Meaning Preservation** – BERTScore between original sentence and de-personified candidate sentence

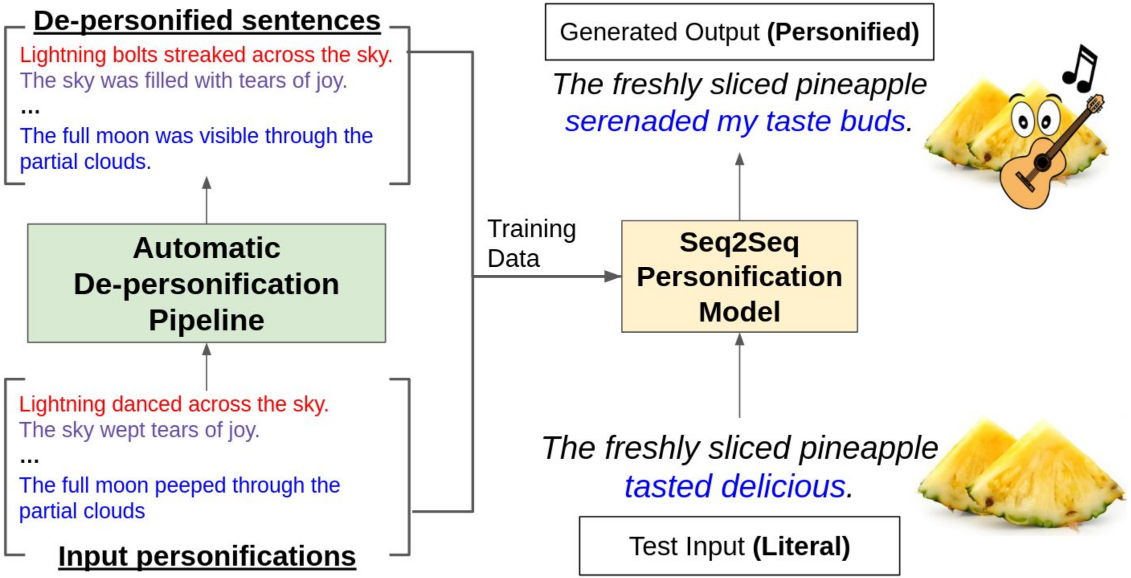
$$S_i = \alpha \cdot (-\log(S_{anim.})) + \beta \cdot S_{flue.} + \gamma \cdot S_{mean.}$$

Select the candidate with the highest S_i score.

Training + Generation

- After de-personifying the dataset, we use the personification+literal pairs to train a seq2seq model with the literal sentences. as the input and personified sentences as outputs.

- Specifically, we use the BART model.



Experimental Setup

Models:

- COMET: No training at all. Adapt our de-personification pipeline, but this time to personification generation. Use IsA(x,"person"), CapableOf(TOPIC, y), and CapableOf("person", y) to generate candidates + rank to select the best personifications.
- Baseline-BART: Similar to COMET, except use BART to generate candidates
- PINEAPPLE-BART: Our proposed model (seq2seq training with personification+literal training pairs)

Evaluation metrics:

- Automatic: BLEU, BERTScore, Fluency, Animacy
- Human (1 to 5 scale): Personificationhood, Appropriateness, Fluency, Interestingness, Meaning Preservation

Results (Automatic Metrics)

- **BLEU** and **BERTScore** – measure if outputs preserve meaning of original
- **Fluency** – generation losses (log-perplexity) using GPT2
- **Animacy** – $\mathcal{A}_{human,\text{ATT}} - \mathcal{A}_{\text{TOPIC,ATT}}$ as previously defined

	BLEU		BERTScore			
	Input	**Gold**	**Input**	**Gold**	**Fluency ↓**	**Animacy**
Human Annotation	0.172	1.000	0.749	1.000	5.264	0.332
COMET	0.127	0.128	0.655	0.569	6.366	-2.028
BL-BART	0.132	0.133	0.728	0.617	**4.573**	0.106
PA-BART	**0.153**	**0.160**	**0.748**	**0.636**	5.460	**0.679**

Table 2: Average automatic evaluation results. The best-scoring method for each metric is highlighted in **bold**. Higher scores are better for all metrics except for fluency.

Results (Human Evaluation)

Human annotators were asked to score each model's outputs on a scale of 1 to 5 along 5 dimensions, as shown below:

	Personificationhood	**Appropriateness**	**Fluency**	**Interestingness**	**Meaning Preservation**
Human Annotation	3.763	4.175	4.138	3.667	3.913
COMET	3.525	3.563	3.738	1.801	3.550
BL-BART	3.500	3.938	**4.188**	2.006	3.750
PA-BART	**3.738**	**4.000**	4.138	**2.782**	**3.875**

Table 3: Average human evaluation results. The best-scoring method for each metric is highlighted in **bold**.

Results (Qualitative Analysis)

Method	Text
Literal Input	You are at a business dinner with your boss when your phone rings out loud (ex.1)
Human Ref	You are at a business dinner with your boss when your phone starts singing out loud
COMET	You are at a business dinner with your boss when your phone beep out loud
BL-BART	You are at a business dinner with your boss when your phone rings and you answer out loud
PA-BART	You are at a business dinner with your boss when your phone yells out loud
Literal Input	In most horror settings, silence is key. (ex.2)
Human Ref	In most horror settings, silence is the protagonist.
COMET	In most horror settings, silence scary.
BL-BART	In most horror settings, silence is preferred.
PA-BART	In most horror settings, silence is a ghost.
Literal Input	Her relationships with family and friends are very difficult (ex.3)
Human Ref	Her relationships with family and friends behave very unusually
COMET	Her relationships with family and friends serious
BL-BART	Her relationships with family and friends have always been strong.
PA-BART	Her relationships with family and friends are very lonely
Literal Input	The sound hit Frank loud enough to make your ear hurt (ex.6)
Human Ref	The sound slapped Frank loud enough to make your ear hurt
COMET	The sound echo Frank loud enough to make your ear sense sound
BL-BART	The sound of Frank Sinatra is loud enough to make your ear ring.
PA-BART	The sound clapped loud enough to make your ear cry

- Can capture cases where the ATTRIBUTE is a noun (“is a ghost”), a verb (“yells”), and an adjective (“very lonely”)

- Can replace and generate multi-word phrases (e.g. “key”→”a ghost”, “hit Frank”→”clapped”)

- Can replace multiple segments in a single sentence (last row: “hit Frank” → ”clapped”, “hurt” → “cry”)

PINEAPPLE: Personifying INanimate Entities by Acquiring Parallel Personification data for Learning Enhanced generation

arXiv preprint:
https://arxiv.org/abs/2209.07752

Dataset + code:
https://github.com/sedrickkeh/PINEAPPLE