Content preserving text generation with attribute controls

Lajanugen Logeswaran, Honglak Lee, Samy Bengio
University of Michigan, Google Brain

March 19, 2019

Outline

1. Introduction
2. Formulation
   - Formulation
   - Content compatibility
   - Attribute compatibility
3. Experiments
   - Metrics
   - Sentiment Experiments
   - Monolingual Translation
   - Ablative study
   - Simultaneous control of multiple attributes
4. Conclusion

Outline

Introduction

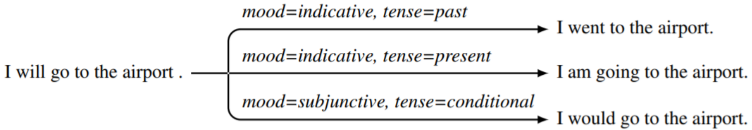
- The style transfer problem which aims to change more abstract properties of an image has seen significant advances.
- The discrete sequential natural of language makes it difficult to approach language problems in a similar manner.
- The focus of this work is on the problem of modifying textual attributes in sentences.
- Create a model that can control multiple attributes of generated text at the same time.

Outline

Formulation

- K attributes of interest $\{a_1, \ldots, a_K\}$.
- A set of labeled sentences $D = \{(x^n, l^n)\}_{n=1}^N$ (l^n is a set of labels for a subset of the attributes)
- Given a sentence x and attributes $l^{'} = (l_1, \ldots, l_K)$, the goal is to produce a sentence that shares the content of x, but reflects the attribute values specified by $l^{'}$.

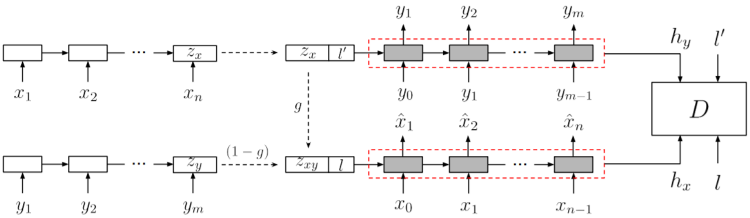
Formulation



Model Overview

- Generative Model $G = (G_{enc}, G_{dec})$
- G should generate a sentence that is closely related in meaning to the input sentence and consistent with the attributes.
- $G_{enc} : z_x = G_{enc}(x)$
- $G_{dec} : y \sim P_G(\cdot|z_x, l')$

Model Overview



Content compatibility

Two types of reconstruction losses to encourage content compatibility.

- Autoencoding loss

$$\mathcal{L}^{ae}(x, l) = -log P_G(x|z_x, l)$$

- Back-translation loss

$$\mathcal{L}^{bt}(x, l) = -log P_G(x|z_y, l)$$

Content compatibility

- A common pitfall of the auto-encoding loss in auto-regressive models.
  - Simply copy the input sequence without capturing and informative features.
- A de-nosing formulation is often considered: deleting, swapping or re-arranging words.
  - However, the generated sample y can be mismatched in content from x

Content compatibility

This paper addresses these issues by interpolating the latent representations of ground truth sentence x and generated sentence y.

- merge the autoencoding and back-translation losses by fusing the two latent representations z_x, z_y

  $$z_{xy} = g \odot z_x + (1 - g) \odot z_y$$

  where g is a binary random vector of values sampled from a Bernoulli distribution.

- Intepolated reconstruction loss

  $$\mathcal{L}^{int} = E_{(x,l) \sim p_{data}, y \sim p_G(\cdot|z_x, l')}[-log p_G(x|z_{xy}, l)]$$

Attribute compatibility

- Adversarial loss

$$\mathcal{L}^{adv} \min_{G} \max_{D} \mathbb{E}[logD(h_x, l) + log(1 - D(h_y, l^{'}))]$$

  It is possible that the discriminator ignores the attributes and makes the real/fake decision based on just the hidden states, or vice versa.

- To prevent this situation, a new objective is proposed

$$\mathcal{L}^{adv} \min_{G} \max_{D} \mathbb{E}[2logD(h_x, l) + log(1 - D(h_y, l^{'})) + log(1 - D(h_x, l^{'}))]$$

- The overall loss function

$$\mathcal{L}^{int} + \mathcal{L}^{adv}$$

Outline

Metrics

- Attribute accuracy
  A pre-trained sentiment classifier
- Content compatibility

$$f_{content}(M, M^{'}) = 0.5[\mathbb{E}_{x \sim D_{src}} BLEU(x, M^{'} \circ M(x)) + \\ \mathbb{E}_{x \sim D_{tgt}} BLEU(x, M \circ M^{'}(x))]$$

  where $M \circ M^{'}(x)$ represents translating $x \in D_{src}$ to domain D_{tgt} and then back to D_{src}.
- Fluency
  A pre-trained language model

Sentiment Experiments

- Data
  Restaurant reviews dataset(447k/128k) & IMDB review corpus(128k/36k)

Sentiment Experiments

- Quantitative evaluation

	Yelp Reviews				IMDB Reviews			
Model	Attribute ↑	Content ↑		Fluency ↓	Attribute ↑	Content ↑		Fluency ↓
	Accuracy	B-1	B-4	Perp.	Accuracy	B-1	B-4	Perp.
Ctrl-gen [18]	76.36%	11.5	0.0	156	76.99%	15.4	0.1	94
Cross-align [22]	90.09%	41.9	3.9	180	88.68%	31.1	1.1	63
Ours	**90.50%**	**53.0**	**7.5**	**133**	**94.46%**	**40.3**	**2.2**	**52**

Sentiment Experiments

- Qualitative evaluation

Restaurant reviews	
negative → positive	
Query	*the people behind the counter were not friendly whatsoever .*
Ctrl gen [18]	the food did n't taste as fresh as it could have been either .
Cross-align [22]	the owners are the staff is so friendly .
Ours	the people at the counter were very friendly and helpful .
positive → negative	
Query	*they do an exceptional job here , the entire staff is professional and accommodating !*
Ctrl gen [18]	very little water just boring ruined !
Cross-align [22]	they do not be back here , the service is so rude and do n't care !
Ours	they do not care about customer service , the staff is rude and unprofessional !
Movie reviews	
negative → positive	
Query	*once again , in this short , there isn't much plot .*
Ctrl gen [18]	it's perfectly executed with some idiotically amazing directing .
Cross-align [22]	but <unk> , , the film is so good , it is .
Ours	first off , in this film , there is nothing more interesting .
positive → negative	
Query	*that's another interesting aspect about the film .*
Ctrl gen [18]	peter was an ordinary guy and had problems we all could <unk> with
Cross-align [22]	it's the <unk> and the plot .
Ours	there's no redeeming qualities about the film .

Sentiment Experiments

- Human evaluation

Model	Attribute	Content	Fluency
Ctrl-gen [18]	66.0%	6.94%	2.51
Cross-align [22]	91.2%	22.04%	2.54
Ours	**92.8%**	**55.10%**	**3.19**

Monolingual Translation

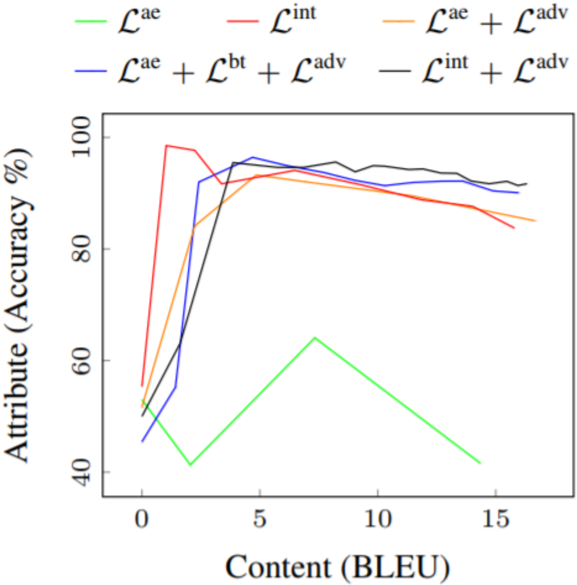
- Use a dataset of Shakespeare plays
- 17k pairs for training and 2k,1k pairs respectively for development and test. All remaining 80k are considered unpaired.
- Train the model using supervised learning and fine-tune on the unpaired data using the proposed objective.

Monolingual Translation

- The results show that the model is capable of finding sentence alignments by exploiting the unlabeled data.

Supervision	Model	BLEU
Paired data	Seq2seq	10.4
	Seq2seq-bi	11.15
Unpaired data	Ours	7.65
Paired + Unpaired data	Ours	**13.89**

Ablative study



Simultaneous control of multiple attributes

Mood	Tense	Voice	Neg.	john was born in the camp
Indicative	Past	Passive	No	john was born in the camp .
Indicative	Past	Passive	Yes	john wasn't born in the camp .
Indicative	Past	Active	No	john had lived in the camp .
Indicative	Past	Active	Yes	john didn't live in the camp .
Indicative	Present	Passive	No	john is born in the camp .
Indicative	Present	Passive	Yes	john isn't born in the camp .
Indicative	Present	Active	No	john has lived in the camp .
Indicative	Present	Active	Yes	john doesn't live in the camp .
Indicative	Future	Passive	No	john will be born in the camp .
Indicative	Future	Passive	Yes	john will not be born in the camp .
Indicative	Future	Active	No	john will live in the camp .
Indicative	Future	Active	Yes	john will not survive in the camp .
Subjunctive	Cond	Passive	No	john could be born in the camp .
Subjunctive	Cond	Passive	Yes	john couldn't live in the camp .
Subjunctive	Cond	Active	No	john could live in the camp .
Subjunctive	Cond	Active	Yes	john couldn't live in the camp .

Outline

Conclusion

- Back-translation is useful for attribute control of discrete data.
- The proposed model can easily extend to the multiple attribute scenario.
- It would be interesting future work to consider attribute with continues values and a much larger set of semantic and syntactic attributes.