Mask & Infill

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han and Songlin Hu

IJCAI 2019

Presented by: Egill Ian Gudmundsson

UNIVERSITY OF WATERLOO

Outline

- Problem definition
- Previous efforts
- Unsolved difficulties
- Mask and Infill method
- Results
- Discussion

Problem Definition

- Style transfer and sentiment transfer usually grouped together
- Take a sentence, change the sentiment (the attribute words) while preserving content
- Sentiment can be, e.g. positive and negative

Positive: Parasite was an incredible film, showing societal issues in a riveting and entertaining manner.

↓

Negative: Parasite was an uninspiring film, showing societal issues in a cliché and trite manner.

Problem Definition

- Corpora *D* only consists of a sentence and its sentiment label (x, s)
- We do not have the same sentence with a different sentiment (x, s')
- Need to implement unsupervised learning since there's no parallel data
- Reconstruction loss is our friend

UNIVERSITY OF WATERLOO

Previous work

- **Delete, Retrieve, Generate (2018)**, J.Li et al
- Used LSTMs to achieve better results
- Picks out the attribute words using saliency scores and deletes them
- Gets a similar sentence from the corpus (using TF-IDF)
- Generates the output sentence by feeding the incomplete sentence and the similar sentence into a neural network
- The saliency method is also called the frequency-ratio method

Previous work

- Delete, Retrieve, Generate (2018), J.Li et al

Parasite was an incredible film, showing societal issues in a riveting and entertaining manner.

It was an uninspiring film, both cliché and trite

Parasite was an film, showing societal issues in a and manner.

Parasite was an uninspiring film, showing societal issues in a cliché and trite manner.

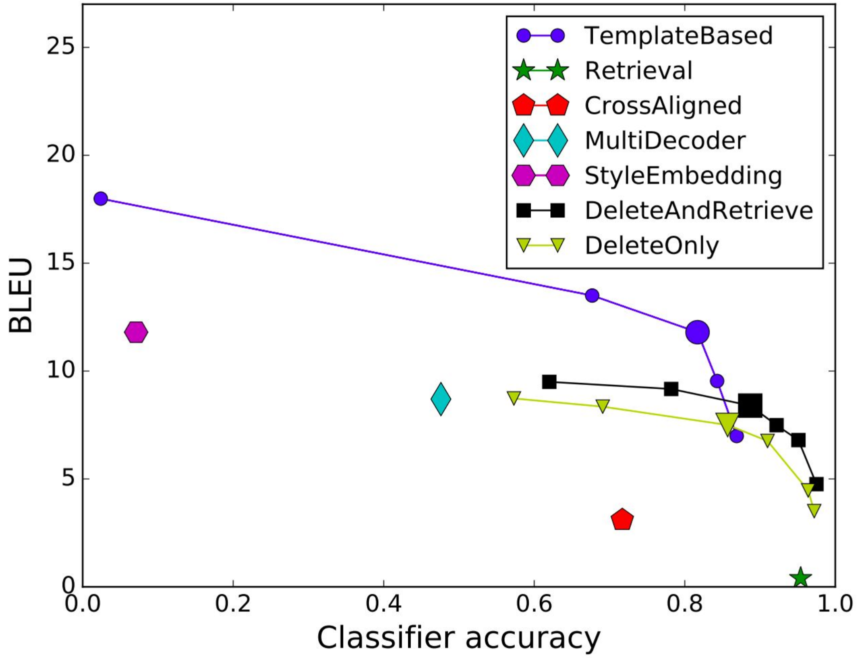
UNIVERSITY OF WATERLOO

Previous work

- **Delete, Retrieve, Generate with Transformers (2019)**, Akhilesh Sudhakar et al
- Swapped the generative network with a Guided Generative Style Transformer and a Blind Generative Style Transformer
- Delete Transformer added in stead of pure saliency score
- Still uses the 3 steps in DRG

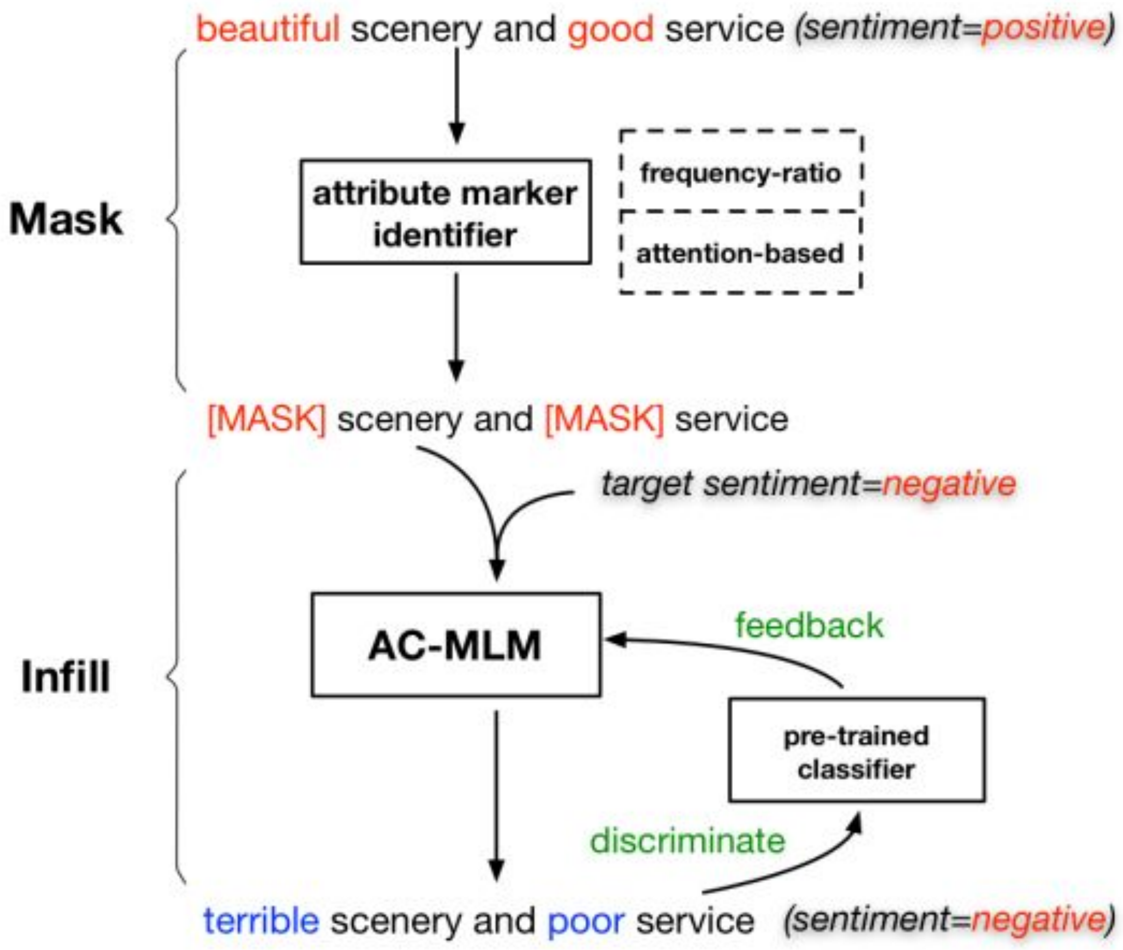
Unsolved difficulties

- Longer sentences still tricky
- May be hard to adapt for multi-sentiment transfer
- Trade-off between content and attributes



Mask & Infill

- Find attribute markers in sentence and mask them
- Generate new attribute markers in the sentence where masked



Masking

- Attribute words identified with a mix of frequency-ratio and attention methods
- Frequency-ratio counts how often an n-gram appears in a sentence with a sentiment

$$s_c(u, a) = \frac{count(u, \mathcal{D}_a) + \lambda}{\left(\sum_{a' \in \mathcal{A}, a' \neq a} count(u, \mathcal{D}_{a'})\right) + \lambda} \quad (1)$$

- If the score is above threshold, marked as attribute marker
- Relies on good corpus, fails if corpus quality is subpar

Masking

- Attribute words identified with a mix of frequency-ratio and attention methods
- Frequency-ratio counts how often an n-gram appears in a sentence with a sentiment

$$s_c(u, a) = \frac{\boxed{count(u, \mathcal{D}_a)} + \lambda}{\boxed{\left(\sum_{a' \in \mathcal{A}, a' \neq a} count(u, \mathcal{D}_{a'})\right)} + \lambda} \quad (1)$$

- If the score is above threshold, marked as attribute marker
- Relies on good corpus, fails if corpus quality is subpar

UNIVERSITY OF WATERLOO

Masking

- Attention method uses bidirectional LSTM trained to extract the extent to which words contribute to sentiment
- If the attention on a word is more than the average in a sentence, it is marked as attribute

$$\mathbf{H} = (h_1, h_2, \cdots, h_N) \quad (2)$$

$$\mathbf{a} = softmax(\mathbf{w} \cdot tanh(\mathbf{W}\mathbf{H}^T)) \quad (3)$$

$$\mathbf{c} = \mathbf{a} \cdot \mathbf{H} \quad (4)$$

$$\mathbf{y} = \mathbf{softmax}(\mathbf{W}' \cdot \mathbf{c}) \quad (5)$$

- Can fail if too many attribute words vying for attention

Masking

- Using a combination of both attention and frequency make up for each other's faults
- Use frequency for saliency score and attention for likelihood of attribute word being selected

$$s(u, a) = s_c(u, a) * p \quad (6)$$

- If the score is above a certain threshold, mark as attribute
- If frequency method fails (too few or too many words) => default to attention method

UNIVERSITY OF WATERLOO

Infill

- Don't have parallel sentences => need to use reconstruction loss

$$\mathcal{L}_{rec} = - \sum_{a \in \mathcal{A}, t_i \in M} logp(t_i | \overline{S}, a) \qquad (7)$$

- Discriminator is used to classify sentence attribute

$$\mathcal{L}_{acc} = -logp(\hat{a} | \hat{S}) \qquad (9)$$

- Combination of the two used for loss

$$min_{\theta} \mathcal{L} = \mathcal{L}_{rec} + \eta \mathcal{L}_{acc} \qquad (10)$$

Infill

- Uses Attribute Conditional Masked Language Model (AC-MLM) for language modelling
- Uses soft-sampling to back-propagate gradients by using approximations of the sampled word vector

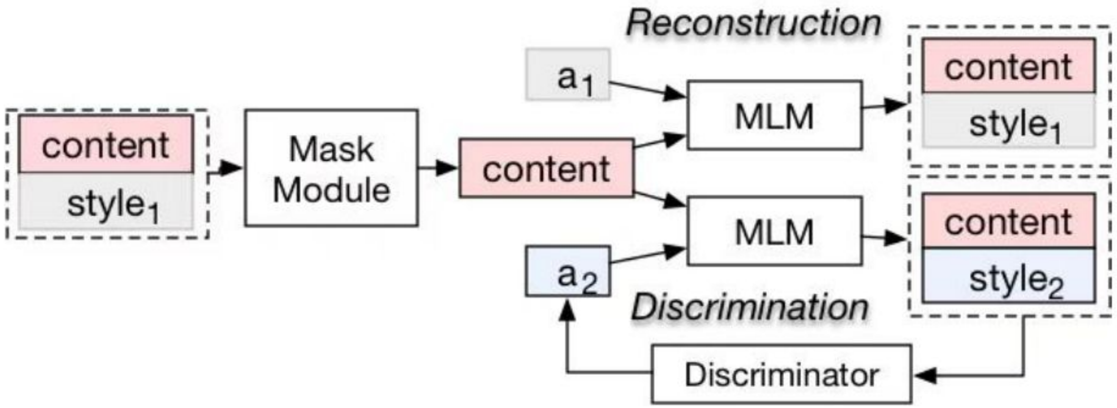
$$\hat{t}_i \sim softmax(\mathbf{o}_t/\tau) \quad (11)$$

UNIVERSITY OF WATERLOO

Mask & Infill Overview

Algorithm 1 Implementation of "Mask and Infill" approach.

Pre-train attention-based classifier *Cls* (Eq.2-5)
Construct attribute marker vocabulary $\mathcal{V}$ (Eq.1,6)
for every sentence S in $\mathcal{D}$ **do**
    Mask attribute markers within S by looking up $\mathcal{V}$, getting $\overline{S}$
    if $\overline{S}$ is too short or $\overline{S}$ is the same as S **then**
        Re-mask with attention weights calculated by *Cls* (Eq.3)
    end if
end for
for each iteration i=1,2,...,M **do**
    Sample a masked sentence $\overline{S}$ with attribute a
    Reconstruct S with $\overline{S}$ and a, calculating $\mathcal{L}_{rec}$ based (Eq.7)
    $\hat{a}$ = the target attribute
    Construct $\hat{S}$ (Eq.8)
    calculating $\mathcal{L}_{acc}$ (Eq.9)
    Update model parameters θ
end for



UNIVERSITY OF WATERLOO

The Boring Stuff

- Model uses BERT_{base} as a language model with original parameters. 12 layers, 12 attention heads, 110 million parameters, 512 token max input
- Segment embedding layer replaced with attribute embedding layer
- Pre-trained discriminator is CNN-based classifier using Word-Piece embeddings
- Balancing parameter and temperature in (10) and (11) respectively are selected via grid-search
- BERT is fine-tuned as AC-MLM for 10 epochs and then 6 epochs are trained with discriminator constraint applied

Results

	YELP			AMAZON		
	Gra	Con	Att	Gra	Con	Att
DeleteAndRetrieval	3.4	3.5	3.6	3.5	3.2	3.3
w/frequency-ratio						
AC-MLM-SS	3.9	3.2	4.2	3.8	3.6	3.7
w/attention-based						
AC-MLM-SS	4.0	3.8	**4.4**	3.9	3.7	3.7
w/fusion-method						
AC-MLM-SS	**4.2**	**4.0**	**4.4**	**4.1**	**4.0**	**4.0**

Table 5: Human evaluation results on two datasets. We show average human ratings for grammaticality (Gra), content preservation (Con), target attribute match (Att).

	YELP		AMAZON	
	ACC (%)	BLEU	ACC(%)	BLEU
CrossAligned	73.7	3.1	74.1	0.4
StyleEmbedding	8.7	11.8	43.3	10.0
MultiDecoder	47.6	7.1	68.3	5.0
CycledReinforce	85.2	9.9	77.3	0.1
TemplateBased	81.7	11.8	68.7	27.1
RetrievalOnly	95.4	0.4	70.3	0.9
DeleteOnly	85.7	7.5	45.6	24.6
DeleteAndRetrieval	88.7	8.4	48.0	22.8
w/frequency-ratio				
AC-MLM	55.0	12.7	28.7	31.0
AC-MLM-SS	90.5	11.6	75.7	26.0
w/attention-based				
AC-MLM	41.5	**15.9**	31.2	**32.1**
AC-MLM-SS	97.3	14.1	75.9	28.5
w/fusion-method				
AC-MLM	43.5	15.3	42.9	30.7
AC-MLM-SS	**97.3**	14.4	**84.5**	28.5

Table 2: Automatic evaluation performed by tools from [Li *et al.*, 2018][5]. "ACC" indicates accuracy, "BLEU" measures content similarity between the outputs and the human references. "AC-MLM", represents attribute conditional masked language model. "w/" represents "with"."-SS" represents with soft-sampling.

Discussion

- Hasn't been tested with multi-sentiment transfer
- Still hasn't fully tackled the content-attribute trade-off

$$min_\theta \mathcal{L} = \mathcal{L}_{rec} + \eta \mathcal{L}_{acc} \qquad (10)$$

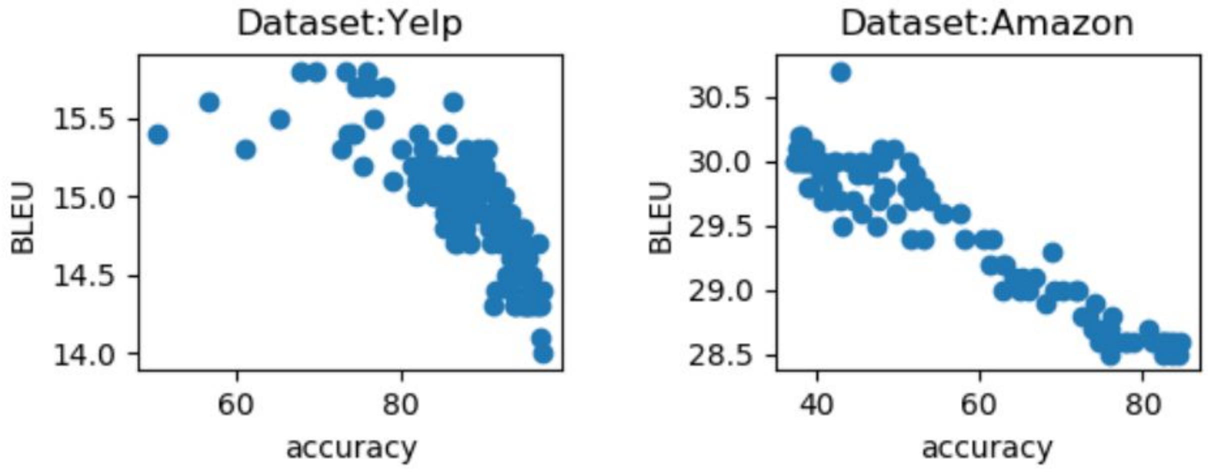


Figure 3: The trend of BLEU with the increase of accuracy.

Discussion

- Can it be used for *style* transfer?

?

Shakespeare

There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy

Lt. Com. Data

Horatio, there are many items that are left unaccounted for in your philosophy

References

- Juncen Li, Robin Jia, He He, Percy Liang. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. In NAACL 2018.
- Akhilesh Sudhakar, Bhargav Upadhyay, Arjun Maheswaran. Transforming Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. In EMNLP 2019.
- Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, Songlin Hu. "Mask and Infill" : Applying Masked Language Model to Sentiment Transfer. In IJCAI 2019.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL 2019.