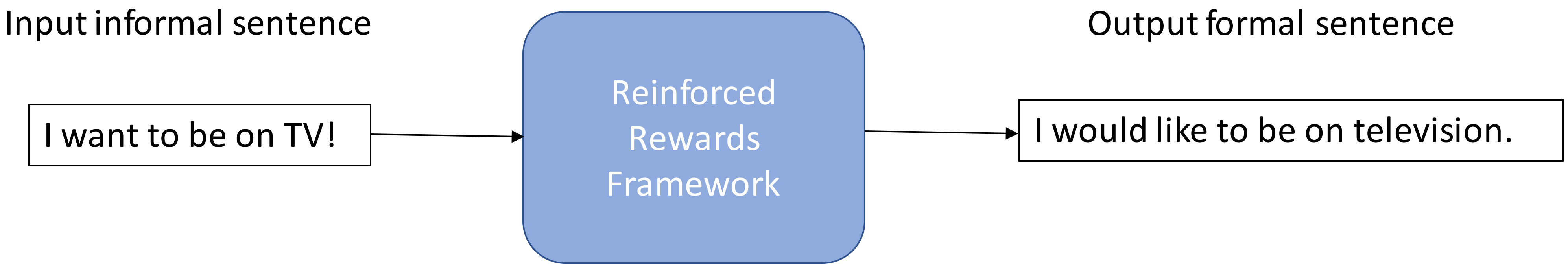
Reinforced Rewards Framework for Text Style Transfer

Abhilasha Sancheti[1,2], Kundan Krishna[3], Balaji Vasan Srinivasan[1] and N. Anandhavelu[1]

1 2 3

Goal: Transfer Style of Text

- Transform style of given text from one form to another
    - Formal to informal (vice-versa)
    - Modern English to Shakespearean English (vice-versa)
    - Exciting to non-exciting (vice-versa)
- Wide variety of applications in content creation



Related Work

- Parallel style transfer
  - Xu et al. 2012[1] introduced a parallel corpora and a phrase-based translation model to modernize Shakespearean English sentences
  - Jhamtani et al. 2017[2] proposed a copy-enriched sequence to sequence model for shakespearizing modern English
  - Rao et al. 2018[3] introduced a parallel corpus of formal and informal sentences

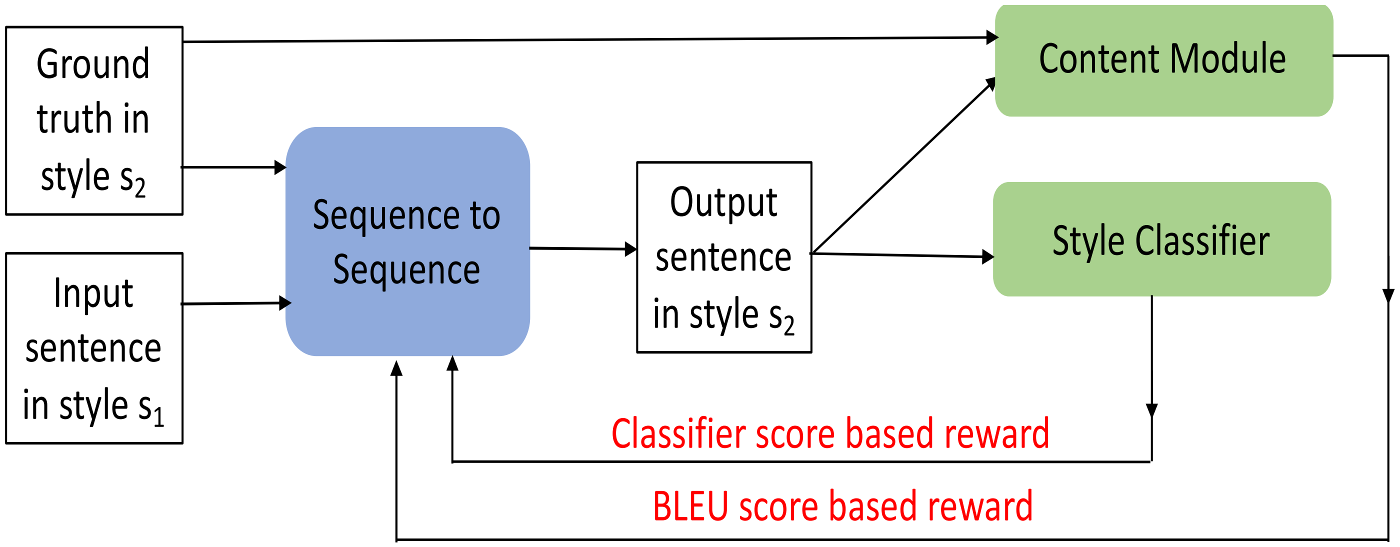
1. Xu, W., Ritter, A., Dolan, B., Grishman, R., Cherry, C.: Paraphrasing for style. Proceedings of COLING 2012 pp. 2899–2914 (2012)
2. Jhamtani, Harsh, et al. "Shakespearizing modern language using copy-enriched sequence-to-sequence models." arXiv preprint arXiv:1707.01161 (2017).
3. Rao, S., & Tetreault, J. (2018). Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. arXiv preprint arXiv:1803.06535.

Related Work

- Non-parallel style transfer
  - Shen et al. 2017[4] assume a shared latent content distribution and propose a method that leverages refined alignment of latent representations
  - Li et al. 2018[5] define style in terms of attributes (such as, sentiment) localized to parts of the sentence and learn to disentangle style from content in an unsupervised setting
- Contributions
  - Sentence level loss terms instead of word level
  - Existing work do not optimize over content preservation and transfer strength metrics but to generate sentences matching reference
  - Reinforced rewards framework

4. Shen, Tianxiao, et al. "Style transfer from non-parallel text by cross-alignment." Advances in neural information processing systems. 2017.
5. Li, J., Jia, R., He, H., Liang, P.: Delete, retrieve, generate: A simple approach to sentiment and style transfer. arXiv preprint arXiv:1804.06437 (2018)

Reinforced Framework



$$P_t(\mathrm{w}) = \delta\, P_t^{RNN}(\mathrm{w}) + (1-\delta)\, P_t^{PTR}(\mathrm{w})$$

$$L_{ml} = -\sum_{t=1}^{m} \log(P_t(y_t*)))$$

1. Jhamtani, Harsh, et al. "Shakespearizing modern language using copy-enriched sequence-to-sequence models." arXiv preprint arXiv:1707.01161 (2017).

Content Module: Rewarding Content Preservation

- Leverage Self-Critic Sequence Training[1] (SCST) to optimize the framework with BLEU score as reward
- BLEU measures the overlap between the ground truth and generated sentence

$$L_{cp} = (\mathrm{r}(y') - \mathrm{r}(y^s)) \sum_{t=1}^{m} \log(p(y_t^s | y_{1:t-1}^s, x))$$

  - y^s is sampled from $p(y_t^s | y_{1:t-1}^s, x)$
  - y' is greedy output
- Note that metric is not required to be differentiable

1. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR. vol. 1, p. 3 (2017)

Style Classifier: Rewarding Transfer Strength

- Formal measure for transfer strength required to use SCST formulation
- Train a CNN-based classifier[1] to predict the likelihood that given sentence belongs to target style
- Likelihood taken as proxy to the reward for transfer strength

$$L_{ts} = \begin{cases} -\log(1 - s(y')), & \text{high to low level} \\ -log(s(y')), & \text{low to high level} \end{cases}$$

- y' is the greedily generated output and $s(y')$ is the likelihood score predicted by the classifier for y'

1. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprintarXiv:1408.5882 (2014

Evaluation

- Three tasks
  - Reinforcing formality (GYAFC dataset)[1]
  - Beyond formality; reinforcing excitement
  - Beyond affective elements (English dataset)[2]
- Metrics
  - Content Preservation: **BLEU** score between model output and ground truth reference
  - Transfer strength: fraction of generated sentences belonging to the target style (**Accuracy**)
  - **Overall**: $\frac{\text{BLEU x Accuracy}}{\text{BLEU+Accuracy}}$

1. Rao, S., Tetreault, J.: Dear sir or madam, may i introduce the gyafc dataset: Corpus, bench-marks and metrics for formality style transfer. In: Proceedings of the 2018 Conference ofthe North American Chapter of the Association for Computational Linguistics: Human Lan-guage Technologies, Volume 1 (Long Papers). vol. 1, pp. 129–140 (2018)
2. Jhamtani, Harsh, et al. "Shakespearizing modern language using copy-enriched sequence-to-sequence models." arXiv preprint arXiv:1707.01161 (2017).

Baselines

- CopyNMT[1]: base model
- Cross-Aligned[2]: unsupervised cross-alignment model
- Transformer[3]: train a transformer-based translation model on style transfer parallel dataset

1 Jhamtani, Harsh, et al. "Shakespearizing modern language using copy-enriched sequence-to-sequence models." arXiv preprint arXiv:1707.01161 (2017).
2. Shen, Tianxiao, et al. "Style transfer from non-parallel text by cross-alignment." Advances in neural information processing systems. 2017.
3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł.,Polosukhin, I.: Attention is all you need. In: Advances in Neural Information ProcessingSystems. pp. 5998–6008 (2017)

Experiments: Reinforcing Formality

- Evaluate our model on GYAFC[1] dataset
  - Parallel corpora for formal-informal text
- Ablation study to demonstrate the improvement in performance of the model with new loss terms
  - CopyNMT: Trained with L_{ml}
  - TS: Trained with L_{ml} followed by $\alpha L_{ml} + \gamma L_{ts}$
  - CP: Trained with L_{ml} followed by $\alpha L_{ml} + \beta L_{cp}$
  - TS+CP: Trained with L_{ml} followed by $\alpha L_{ml} + \beta L_{cp} + \gamma L_{ts}$
  - TS→CP: Trained with L_{ml} followed by $\alpha L_{ml} + \gamma L_{ts}$ and finally with $\alpha L_{ml} + \beta L_{cp}$
  - CP→TS: Trained with L_{ml} followed by $\alpha L_{ml} + \beta L_{cp}$ and finally with $\alpha L_{ml} + \gamma L_{ts}$

1. Rao, S., & Tetreault, J. (2018). Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. arXiv preprint arXiv:1803.06535.

Results: Ablation Study

	Informal to Formal			Formal to Informal		
Models	BLEU↑	Accuracy↑	Overall↑	BLEU↑	Accuracy↑	Overall↑
CopyNMT	0.263	0.774	0.196	0.280	0.503	0.180
TS	0.240	0.801	0.184	0.271	0.527	0.179
CP	0.272	0.749	0.199	0.281	0.487	0.178
TS+CP	0.259	0.772	0.194	0.271	0.527	0.179
CP→TS	0.227	**0.817**	0.178	0.259	**0.5441**	0.175
TS→CP	**0.286**	0.723	**0.205**	**0.298**	0.516	**0.189**

Results: Ablation Study

	Informal to Formal			Formal to Informal		
Models	BLEU↑	Accuracy↑	Overall↑	BLEU↑	Accuracy↑	Overall↑
CopyNMT	0.263	0.774	0.196	0.280	0.503	0.180
TS	0.240	0.801	0.184	0.271	0.527	0.179
CP	0.272	0.749	0.199	0.281	0.487	0.178
TS+CP	0.259	0.772	0.194	0.271	0.527	0.179
CP→TS	0.227	**0.817**	0.178	0.259	**0.5441**	0.175
TS→CP	**0.286**	0.723	**0.205**	**0.298**	0.516	**0.189**

Results: Ablation Study

	Informal to Formal			Formal to Informal		
Models	BLEU↑	Accuracy↑	Overall↑	BLEU↑	Accuracy↑	Overall↑
CopyNMT	0.263	0.774	0.196	0.280	0.503	0.180
TS	0.240	0.801	0.184	0.271	0.527	0.179
CP	0.272	0.749	0.199	0.281	0.487	0.178
TS+CP	0.259	0.772	0.194	0.271	0.527	0.179
CP→TS	0.227	**0.817**	0.178	0.259	**0.5441**	0.175
TS→CP	**0.286**	0.723	**0.205**	**0.298**	0.516	**0.189**

Results: Ablation Study

	Informal to Formal			Formal to Informal		
Models	BLEU↑	Accuracy↑	Overall↑	BLEU↑	Accuracy↑	Overall↑
CopyNMT	0.263	0.774	0.196	0.280	0.503	0.180
TS	0.240	0.801	0.184	0.271	0.527	0.179
CP	0.272	0.749	0.199	0.281	0.487	0.178
TS+CP	0.259	0.772	0.194	0.271	0.527	0.179
CP→TS	0.227	**0.817**	0.178	0.259	**0.5441**	0.175
TS→CP	**0.286**	0.723	**0.205**	**0.298**	0.516	**0.189**

Results: Formality Dataset

	Informal to Formal			Formal to Informal		
Models	**BLEU↑**	**Accuracy↑**	**Overall↑**	**BLEU↑**	**Accuracy↑**	**Overall↑**
Transformer	0.125	**0.933**	0.110	0.099	**0.894**	0.089
Cross-Aligned	0.116	0.670	0.098	0.117	0.766	0.101
CopyNMT	0.263	0.774	0.196	0.280	0.503	0.180
TS→CP (Proposed)	**0.286**	0.723	**0.205**	**0.298**	0.516	**0.189**

Experiments: Beyond Formality

- Evaluate on Excitement dataset to demonstrate generalizability
- Curated this dataset using reviews from Yelp[1]
- Reviews with rating >= 3 considered as exciting sentences
- Asked Amazon Mechanical Turkers to rewrite the exciting sentences to make them sound boring or non-exciting
- Asked AMT to rate rewrites and given sentences on a Likert scale of 1(no excitement) to 5 (very high excitement)

1. https://www.yelp.com/dataset

Results: Beyond Formality

	Exciting to Non-exciting			Non-exciting to Exciting		
Models	**BLEU↑**	**Accuracy↑**	**Overall↑**	**BLEU↑**	**Accuracy↑**	**Overall↑**
Transformer	0.077	**0.922**	0.071	0.069	0.605	0.062
Cross-Aligned	0.059	0.818	0.055	0.061	0.547	0.054
CopyNMT	0.143	0.919	0.124	0.071	**0.813**	0.065
TS→CP (Proposed)	**0.153**	**0.922**	**0.131**	**0.088**	0.744	**0.078**

Experiments: Beyond Affective Elements

- Evaluate our model on modern English and Shakespearean English dataset[1]

	Modern to Shakespearean			Shakespearean to Modern		
Models	**BLEU↑**	**Accuracy↑**	**Overall↑**	**BLEU↑**	**Accuracy↑**	**Overall↑**
Transformer	0.027	**0.736**	0.026	0.046	**0.915**	0.043
Cross-Aligned	0.044	0.614	0.041	0.049	0.537	0.044
CopyNMT	0.104	0.495	0.085	0.111	0.596	0.093
TS→CP (Proposed)	**0.127**	0.489	**0.100**	**0.137**	0.567	**0.110**

1. Jhamtani, Harsh, et al. "Shakespearizing modern language using copy-enriched sequence-to-sequence models." arXiv preprint arXiv:1707.01161 (2017).

Human Evaluation

- Ask AMT to rate model outputs and reference
- 3 annotators per output
- Content Preservation: Likert scale of 6

  6: Completely equivalent, 5: Mostly equivalent, 4: Roughly equivalent, 3: Not equivalent but share some details, 2: Not equivalent but on same topic,

  1: Completely dissimilar
- Transfer Strength: Likert scale of 5

  5: Very Informal (Very high excitement)

  1: Very formal (No excitement at all)

Results: Human Evaluation

Task	Transfer Strength			Content Preservation		
	R>C	R>T	R>S	R>C	R>T	R>S
I-F	88.67	81.34	70.00	70.00	72.67	83.67
F-I	73.34	88.67	61.22	59.34	79.34	91.80
E-NE	64.00	79.34	68.00	60.67	71.34	71.73
NE-E	76.67	70.67	68.00	69.34	74.00	70.00

Table 3: Human evaluation results of 50 randomly selected model outputs. The values represent the % of times annotators rated model outputs from TS→CP (R) as better than the baseline CopyNMT (C), Transformer (T) and Cross-Aligned (S) over the metrics. I-F (E-NE) refers to informal to formal (exciting to non-exciting) task.

Takeaway

- Explicit optimization over metrics helps in boosting the performance
- Generalized approach; works for a variety of style transfer tasks
- Trade-off between content preservation and transfer strength
- As a future work, we intend to study transfer of multiple styles simultaneously

THANKS!

Contact: sancheti@cs.umd.edu