Unsupervised Evaluation Metrics and Learning Criteria for Non-Parallel Textual Transfer

Richard Yuanzhe Pang
New York University
Work done at the University of Chicago and TTIC

Kevin Gimpel
Toyota Technological Institute at Chicago (TTIC)

1 Task and Motivation

X_0, X_1: Two non-parallel corpora of different "styles"
$\mathbf{x}_t^{(i)}$: ith sentence of style t
$\mathbf{y}_t$: style vector for style t
$\mathbf{z}_t^{(i)}$: content vec for ith sent of style t

$E : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z} \quad G : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$

Want $\widetilde{\mathbf{x}}_t^{(i)} = G(\mathbf{y}_{1-t}, E(\mathbf{x}_t^{(i)}, \mathbf{y}_t))$ $\longleftrightarrow$ $x_0, y_0 \xrightarrow{E} z_0 \xrightarrow{y_1} \widetilde{x}_0$

Application Generating textual paraphrases with modified attributes or stylistic properties (politeness, formality, etc.), benefiting dialogue, writing assistance, etc.; See Pang (2019) for more applications

Lack of parallel corpora => Need **unsup learning criteria** and **unsup evaluation metrics**

Three goals **Correct transfer** (by classifier), **semantic similarity**, **fluency**

Datasets Yelp (positive vs. negative), Literature (Dickens vs. Modern)

2 Eval by Transfer Style Accuracy

(1) Acc (post-transfer accuracy) How often was a pretrained classifier convinced of transfer?

INSUFFICIENT!

#ep of training	Acc (of the entire transferred set)	Sim (of the entire transferred set)	Sentence (negative -> positive)
original input			the host that walked us to the table and left without a word .
0.5	0.87	0.65	the food is the best and the food is the .
3.3	0.72	0.75	the owner that went to to the table and made a smile .
7.5	0.58	0.81	The host that walked through the table and are quite perfect !

Above table: Trained using Shen et al. (2017)

3 Improvements to Eval Metrics

(2) Sim (semantic similarity)

Def (i) Embed sentences by avg word embeddings (GloVe, 300d) weighted by idf; (ii) Sim is the avg of the cos sim over all original/transferred sentence pairs

- Also tried METEOR (large Spearman's correlation with Sim)
- Simplicity => efficient & good for widespread adoption

(3) PP (fluency)

Def Measured by language model trained on concat of two corpora

- PP is distinct from fluency, but correlated
- Punished abnormally small PP below

(1+2+3) Summarizing Acc, Sim, PP into one single number called GM

$$GM_{\mathbf{t}}(q) = \left([100 \cdot \text{Acc} - t_1]_+ \cdot [100 \cdot \text{Sim} - t_2]_+ \cdot \min\{[t_3 - \text{PP}]_+, [\text{PP} - t_4]_+\}\right)^{\frac{1}{3}}$$

- Sampled 300 pairs of *transferred* sentences and asked annotators which one is better
- Training params in GM: t's are trained by

$$L_{\text{GM}}(\mathbf{t}) = \max(0, -\text{GM}_{\mathbf{t}}(y^+) + \text{GM}_{\mathbf{t}}(y^-) + 1)$$

- t = (63, 71, 97, −37) in our experiments

4 Learning Criteria

Built on Shen et al. (2017); Encoder-decoder network

Reconstruction loss $x_0, y_0 \xrightarrow{E} z_0 \xrightarrow{y_0} \widehat{x}_0 \overset{\text{loss}}{\longleftrightarrow} x_0$

Adversarial loss D_0 distinguishes b/w $\mathbf{x_0}$ and $\widetilde{\mathbf{x}}_1$, D_1 b/w $\mathbf{x_1}$ and $\widetilde{\mathbf{x}}_0$

Cycle consistency $x_0, y_0 \xrightarrow{E} z_0 \xrightarrow{y_1} \widetilde{x}_0 \xrightarrow{E} \widetilde{z}_0 \xrightarrow{y_0} \widetilde{\widetilde{x}}_0 \overset{\text{loss}}{\longleftrightarrow} x_0$

Paraphrase loss $u, y_0 \xrightarrow{E} z_0 \xrightarrow{y_0} \widehat{u} \overset{\text{loss}}{\longleftrightarrow} v$

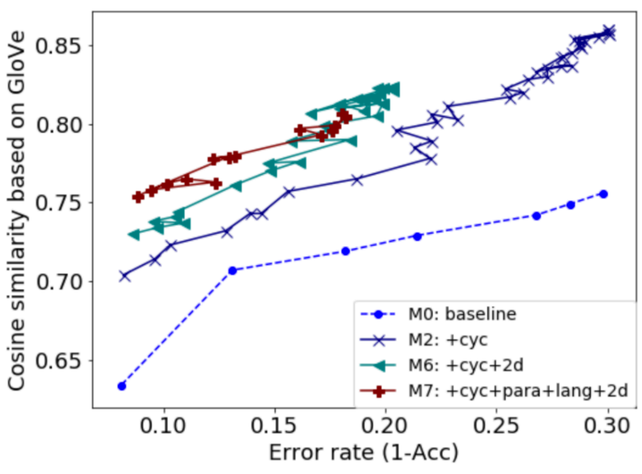
Language model loss $x_0, y_0 \xrightarrow{E} z_0 \xrightarrow{y_1} \widetilde{x}_0 \overset{\text{loss}}{\longleftrightarrow}$ "LM_1"

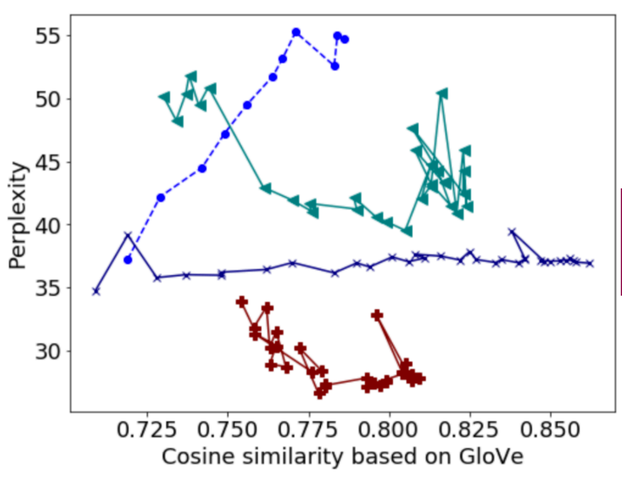
Two sets of discriminators

D_0, D_1 (adv. loss) and D_0', D_1' (WGAN adv. loss)

Model	Losses	Model	Losses
M0	**Shen et al. (2017): rec+adv**	M4	M0+cyc+para
M1	M0+para	M5	M0+cyc+para+lang
M2	**M0+cyc**	**M6**	**M0+cyc+2d**
M3	M0+lang	**M7**	**M0+cyc+para+lang+2d**

5 Result (a): Metric Relationships





Negative correlation b/w Sim and Acc (Generally) positive correlation b/w PP and Sim

6 Result (b): System-Level Validation

Dataset	Models A	Models B	Transfer quality A>B	Transfer quality B>A	Transfer quality Tie	Semantic preservation A>B	Semantic preservation B>A	Semantic preservation Tie	Semantic preservation Δ_{Sim}	Fluency A>B	Fluency B>A	Fluency Tie	Fluency Δ_{PP}
Yelp	M0	M2	9.0	6.0	85.1	1.5	**25.4**	73.1	-0.05	10.4	**23.9**	65.7	0.9
	M0	M7	9.6	14.7	75.8	2.5	**54.5**	42.9	-0.09	4.6	**39.4**	56.1	8.3
	M6	M7	13.7	11.6	74.7	16.0	16.7	67.4	0.01	10.3	20.0	69.7	14.3
	M2	M7	5.8	9.3	84.9	8.1	**25.6**	66.3	-0.04	14.0	**26.7**	59.3	7.4
Literature	M2	M6	4.2	6.7	89.2	16.7	20.8	62.5	0.01	**40.8**	13.3	45.8	-13.3
	M6	M7	15.8	13.3	70.8	**25.0**	9.2	65.8	0.03	14.2	20.8	65.0	14.2

Above table: **Human judgments** b/w transferred sentences from model A and model B

Summary Human judgments in line with automatic measures for semantic preservation and fluency

7 Result (c): Sentence-Level Validation of Metrics

Metric	Method of validation	Yelp	Lit.
Acc	% of machine and human judgments that match	94	84
Sim	Spearman's corr b/w Sim and human ratings of semantic preservation	0.79	0.75
PP	Spearman's corr b/w negative PP and human ratings of fluency	0.81	0.69

- Sampled (from different models) 100 examples each dataset to validate Acc, 150 examples for Sim and PP
- Human ratings of Sim and PP: On a scale of 1 to 4

8 Examples

Model	GM	Sentence	Style
Original	—	the mozzarella sub is absolutely amazing .	Positive
M0	10.0	the front came is not much better .	Negative
M7	22.8	the cheese sandwich is absolutely awful .	Negative
Original	—	they are completely unprofessional and have no experience .	Negative
M0	10.0	they are super fresh and well !	Positive
M7	22.8	they are very professional and have great service .	Positive
Original	—	i declined on their offer , but appreciated the gesture !	Positive
M0	10.0	i asked on their reviews , they are the same time !	Negative
M7	22.8	i paid for the refund , and explained the frustration !	Negative
Original	—	i conjure you , tell me what is the matter .	Dickens
M0	8.81	i 'm sorry , i 'm sure i 'm going to be , but i was a little man .	Modern
M2	12.8	i 'm telling you , tell me what 's the time .	Modern
M6	12.8	i am telling you , tell me what 's the matter .	Modern
Original	—	it whispered to me about my new strength and abilities .	Modern
M0	8.81	it is not a little man .	Dickens
M2	12.8	it appears to me about my new strength and desire .	Dickens
M6	12.8	it appears to me my new strength and desire .	Dickens

Textual transfer evaluation + model code: yzpang.me

Bibliography

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: exploration and evaluation. *In 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *In Advances in Neural Information Processing Systems*, pages 2672–2680.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics.

Richard Yuanzhe Pang. 2019. The Daunting Task of Real-World Textual Style Transfer Auto-Evaluation. *arXiv preprint arXiv:1910.03747*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *In Advances in Neural Information Processing Systems 30*, pages 6833–6844. Curran Associates, Inc.