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

1 Textual Transfer Model

1.1 Summary

We iteratively update (1) θ_{D_0}, θ_{D_1}, $\theta_{D'_0}$, and $\theta_{D'_1}$ by gradient descent on L_{adv_0}, L_{adv_1}, $L_{adv'_0}$, and $L_{adv'_1}$, respectively, and (2) θ_E, θ_G by gradient descent on $L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{para} + \lambda_3 L_{cyc} + \lambda_4 L_{lang} - \lambda_5(L_{adv_0} + L_{adv_1}) - \lambda_6(L_{adv'_0} + L_{adv'_1})$.

1.2 Full Algorithm

Please refer to Algorithm 1.

2 Tables and Plots in Results

Figures 2a and 2b show the learning trajectories for the Literature dataset, which show similar trends as those for Yelp. While the plots for the two datasets appear different from an initial glance, comparing similarities at fixed error rates and comparing perplexities at fixed similarities reveals that the results largely resemble those for the Yelp dataset. The baseline M0 struggles on the Literature dataset. The particularly low perplexity for M0 does not indicate fluent sentences, but rather the piecing together of extremely common words and phrases.

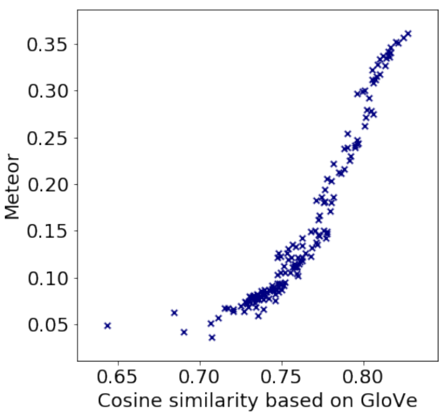


Figure 1: Met by Sim using the Literature dataset

Algorithm 1: Training procedure

Pretrain language models LM_0 and LM_1 to be used in language modeling loss L_{lang}.
Initialize parameters $(\theta_E, \theta_G, \theta_{D_0}, \theta_{D_1}, \theta_{D'_0}, \theta_{D'_1})$.
while *losses have not converged* **do**
- Sample mini-batch $\{\mathbf{x}_t^{(i)}\}_{i=1}^k$ from $\mathbf{X}_t$, and obtain transferred sentences $\{\widetilde{\mathbf{x}}_t^{(i)}\}_{i=1}^k$ by running the decoder $G(\mathbf{y}_{1-t}, E(\mathbf{x}_t, \mathbf{y}_t))$, for $t = 0, 1$.
- Get content representations $\mathbf{z}_t^{(i)} = E(\mathbf{x}_t^{(i)}, \mathbf{y}_t)$, and $\widetilde{\mathbf{z}}_t^{(i)} = E(\widetilde{\mathbf{x}}_t^{(i)}, \mathbf{y}_{1-t})$ for $t = 0, 1, \forall i$, where we use $\mathbf{x}_t^{(i)}$ as inputs for the RNNs and $\mathbf{y}_{1-t}$ as initial hidden states for the RNNs.
- Obtain probability distribution of the back-transferred sentences $\{\widetilde{\widetilde{\mathbf{x}}}_t^{(i)}\}_{i=1}^k$ through decoder $G(\mathbf{y}_t, E(\widetilde{\mathbf{x}}_t, \mathbf{y}_{1-t}))$, for $t = 0, 1, \forall i$.
- Unfold G from $(\mathbf{y}_t, \mathbf{z}_t^{(i)})$ (i.e., by using $(\mathbf{y}_t, \mathbf{z}_t^{(i)})$ as initial hidden state of the RNN), and feed in $\mathbf{x}_t^{(i)}$ to obtain $\mathbf{h}_t^{(i)}$; and unfold G from $(\mathbf{y}_{1-t}, \mathbf{z}_t^{(i)})$, and feed in previous output probability distributions to obtain $\widetilde{\mathbf{h}}_t^{(i)}$. This step is done for $t = 0, 1, \forall i$.
- Compute L_{rec} by (1); Compute L_{adv_0} and L_{adv_1} of the first discriminator by (2), and $L_{adv'_0}$ and $L_{adv'_1}$ of the second discriminator by (6); Compute L_{cyc} by (3); Compute L_{para} by (4); Compute L_{lang} by (5).
- Update $\theta_{D_0}, \theta_{D_1}, \theta_{D'_0}$, and $\theta_{D'_1}$ by gradient descent on $L_{adv_0}, L_{adv_1}, L_{adv'_0}$, and $L_{adv'_1}$, respectively.
- Update θ_E, θ_G by gradient descent on $L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{para} + \lambda_3 L_{cyc} + \lambda_4 L_{lang} - \lambda_5(L_{adv_0} + L_{adv_1}) - \lambda_6(L_{adv'_0} + L_{adv'_1})$.

end

In our analysis, we used Sim as the primary metric for semantic preservation. However, if we were to use Met instead (where Met is computed by METEOR scores between original sen-

Yelp	Acc ≈ 0.800					Sim ≈ 0.800				
	Acc(↑)	Sim(↑)	Met(↑)	PP(↓)	GM(↑)	Acc	Sim	Met	PP	GM
M0: Shen et al. (2017)	0.818	0.719	0.165	37.3	10.0	0.591	0.793	0.305	56.1	0.00
M1: M0+*para*	0.819	0.734	0.196	26.3	14.2	0.704	0.798	0.288	31.0	16.3
M2: M0+*cyc*	0.813	0.770	0.271	36.4	18.8	0.795	0.801	0.312	37.4	20.8
M3: M0+*cyc*+*lang*	0.807	0.796	0.257	28.4	21.5	0.792	0.802	0.272	28.7	21.4
M4: M0+*cyc*+*para*	0.798	0.783	0.275	39.7	19.2	0.794	0.799	0.320	39.4	20.3
M5: M0+*cyc*+*para*+*lang*	0.804	0.785	0.254	27.1	20.3	0.781	0.794	0.288	28.0	20.2
M6: M0+*cyc*+*2d*	0.805	**0.817**	**0.322**	43.3	21.6	0.834	0.807	0.321	47.7	21.4
M7: M0+*cyc*+*para*+*lang*+*2d*	0.818	0.805	0.288	29.0	**22.8**	0.830	0.799	0.281	**27.8**	**22.6**

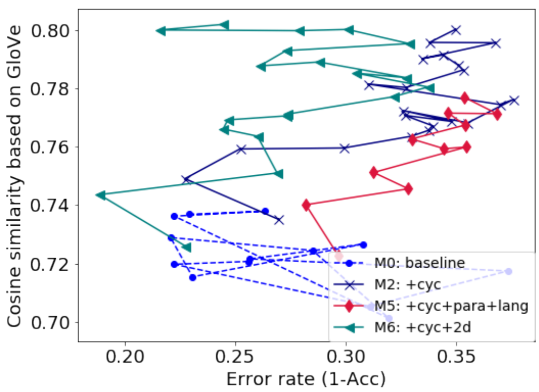
Literature	Acc ≈ 0.700					Sim ≈ 0.750				
	Acc	Sim	Met	PP	GM	Acc	Sim	Met	PP	GM
M0: Shen et al. (2017)	0.694	0.728	0.080	22.3	8.81	n/a	n/a	n/a	n/a	n/a
M1: M0+*para*	0.702	0.747	0.108	23.6	11.7	0.678	0.749	0.106	30.8	10.7
M2: M0+*cyc*	0.692	0.781	0.194	49.9	**12.8**	0.778	0.754	0.109	55.0	14.0
M3: M0+*cyc*+*lang*	0.698	0.754	0.089	39.2	12.0	0.698	0.754	0.089	39.2	12.0
M4: M0+*cyc*+*para*	0.702	0.757	0.117	33.9	**12.8**	0.719	0.756	0.112	29.7	14.0
M5: M0+*cyc*+*para*+*lang*	0.688	0.753	0.089	28.6	11.8	0.727	0.750	0.080	**28.6**	13.7
M6: M0+*cyc*+*2d*	0.704	**0.794**	**0.274**	63.2	**12.8**	0.775	0.758	0.115	55.1	**14.3**
M7: M0+*cyc*+*para*+*lang*+*2d*	0.706	0.768	0.142	49.0	**12.8**	0.749	0.756	0.121	45.6	14.1

Table 1: Results at fixed levels of post-transfer classification accuracy (Acc) and semantic similarity (Sim). Under similar Acc, the best Sim and Met are in bold. Under similar Sim, the best PP is in bold. In both tables, the best GM scores are also in bold. Here, *para* = paraphrase loss, *cyc* = cyclic loss, *lang* = language modeling loss, and *2d* = two pairs of discriminators. Cells with n/a indicate that the model never reaches the corresponding Acc or Sim.

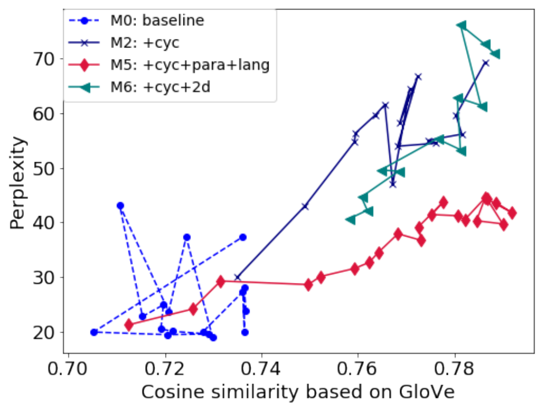
tence and transferred sentence, averaged over sentence pairs), the plots and our conclusions would be largely unchanged. Using the Literature dataset as an example, Figure 1 shows that the correlation between Met and Sim is very large. Specifically, we randomly sample 200 transferred corpora generated using different models, and generated at different times during training. We obtain Met and Sim of each of these 200 transferred corpora using techniques discussed in the main text. We thus have 200 data points, as shown in Figure 1.

3 Examples

Table 2 provides examples of textual transfer.



(a) Cosine similarity (Sim) by error rate ($1 - \mathrm{Acc}$) for Literature.



(b) Perplexity (PP) by cosine similarity (Sim) for Literature.

Figure 2: Learning trajectories with selected models from Table 2 of main text. Metrics are computed on the development sets.

Model	Acc	Sim	PP	GM	Sentence	Style
Original	—	—	—	—	i got my car back and was extremely unhappy .	Negative
M0	0.818	0.719	37.3	10.0	i got my favorite loves and was delicious .	Positive
M7	0.818	0.805	29.0	22.8	i got my car back and was very happy .	Positive
Original	—	—	—	—	the mozzarella sub is absolutely amazing .	Positive
M0	0.818	0.719	37.3	10.0	the front came is not much better .	Negative
M7	0.818	0.805	29.0	22.8	the cheese sandwich is absolutely awful .	Negative
Original	—	—	—	—	they are completely unprofessional and have no experience .	Negative
M0	0.818	0.719	37.3	10.0	they are super fresh and well !	Positive
M7	0.818	0.805	29.0	22.8	they are very professional and have great service .	Positive
Original	—	—	—	—	i would honestly give this place zero stars if i could .	Negative
M0	0.818	0.719	37.3	10.0	i would recommend give this place from everyone again .	Positive
M7	0.818	0.805	29.0	22.8	i would definitely recommend this place all stars if i could .	Positive
Original	—	—	—	—	for all those reasons , we wo n't go back .	Negative
M0	0.818	0.719	37.3	10.0	for all of pizza , you do you go .	Positive
M7	0.818	0.805	29.0	22.8	for all those reviews , i highly recommend to go back .	Positive
Original	—	—	—	—	the owner was super nice and welcoming .	Positive
M0	0.818	0.719	37.3	10.0	the server was extremely bland with all .	Negative
M7	0.818	0.805	29.0	22.8	the owner was very rude and unfriendly .	Negative
Original	—	—	—	—	this is one of the best hidden gems in phoenix .	Positive
M0	0.818	0.719	37.3	10.0	this is one of the worst _num_ restaurants in my life .	Negative
M7	0.818	0.805	29.0	22.8	this is one of the worst restaurants in phoenix .	Negative
Original	—	—	—	—	i declined on their offer , but appreciated the gesture !	Positive
M0	0.818	0.719	37.3	10.0	i asked on their reviews , they are the same time !	Negative
M7	0.818	0.805	29.0	22.8	i paid for the refund , and explained the frustration !	Negative
Original	—	—	—	—	it was a most extraordinary circumstance .	Dickens
M0	0.694	0.728	22.3	8.81	it was a little deal of the world .	Modern
M2	0.692	0.781	49.9	12.8	it was a huge thing on the place .	Modern
M6	0.704	0.794	63.2	12.8	it was a most important effort over the relationship .	Modern
Original	—	—	—	—	i conjure you , tell me what is the matter .	Dickens
M0	0.694	0.728	22.3	8.81	i 'm sorry , i 'm sure i 'm going to be , but i was a little man .	Modern
M2	0.692	0.781	49.9	12.8	i 'm telling you , tell me what 's the time .	Modern
M6	0.704	0.794	63.2	12.8	i am telling you , tell me what 's the matter .	Modern
Original	—	—	—	—	a public table is laid in a very handsome hall for breakfast , and for dinner , and for supper .	Dickens
M0	0.694	0.728	22.3	8.81	the other of the man was a little , and then , and -person- 's eyes , and then -person- .	Modern
M2	0.692	0.781	49.9	12.8	a little table is standing there for all , and for me , and for you .	Modern
M6	0.704	0.794	63.2	12.8	a small table is placed in a very blue room for breakfast , and for dinner , and for dinner .	Modern
Original	—	—	—	—	does n't she know it 's dangerous for a young woman to go off by herself ?	Modern
M0	0.694	0.728	22.3	8.81	do n't have been a little of a man of your own ?	Dickens
M2	0.692	0.781	49.9	12.8	it n't she know it 's dangerous for a little woman to go out from us ?	Dickens
M6	0.704	0.794	63.2	12.8	does n't she know it 's a dangerous act for a young lady to go off by herself ?	Dickens
Original	—	—	—	—	it whispered to me about my new strength and abilities .	Modern
M0	0.694	0.728	22.3	8.81	it is not a little man .	Dickens
M2	0.692	0.781	49.9	12.8	it appears to me about my new strength and desire .	Dickens
M6	0.704	0.794	63.2	12.8	it appears to me my new strength and desire .	Dickens

Table 2: Textual transfer examples

References

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.