MULTIPLE-ATTRIBUTE TEXT STYLE TRANSFER

Motivation

- Text style transfer works well when parallel data is available
    - Performance drops when parallel data is scarce
- The creation of parallel data is difficult, and costly
- Sparsity of datasets of parallel sentences written in a different style
- The discrete nature of the sentence generation process makes it difficult to apply to text techniques such as cycle consistency or adversarial training.

Aim of Paper

- A deeper understanding of the necessary components of style transfer through extensive experiments
- Introduce a generic model based on mixing a denoising auto-encoding loss, replacing the adversarial term with a online back-translation technique and a novel neural architecture combining a pooling operator and support for multiple attributes,
- A new, more challenging and realistic version of existing benchmarks which uses full reviews and multiple attributes per review, as well as a comparison of our approach w.r.t. baselines using both new metrics and human evaluations

Generic Problem

- Given texts from two different domains X and Y (where X could be the domain of poetry and Y the domain of product reviews), and the task is to learn two mappings $F : X \rightarrow Y$ and $G : Y \rightarrow X$, without supervision, i.e., just based on text sampled from the two domains

The task

- Training set = $\mathfrak{D} = (x^i, y^i)_{i \in [1,n]}$

  $x^i \in$ X paired with attribute values in y^i

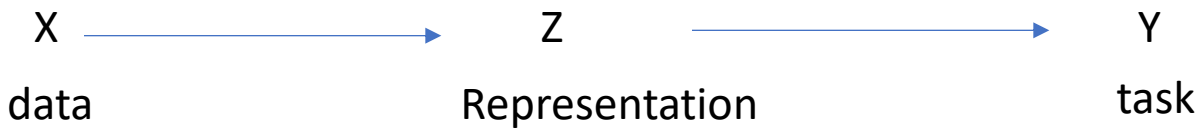
  $y \in$ Y; $y = (y1, \ldots ym)$

  Each attribute $yk \in$ Yk e.g Yk = {bad, neutral, good}

Task: F : X * Y → X that maps (x, ỹ) of an input sentence (x has y attributes originally) and a new set of m attribute values ỹ to a new sentence $\grave{x}$ that has the specified attribute ỹ while retaining as much as possible of the original content of x. Content is anything in x which does not depend on the attributes

Disentanglement

- Create a new representation of data **sufficient** for task
- Remove nuisance information not necessary for the task
- Z = data – nuisance (irrelevant information)



- Z contains **minimal** information
- Z is **invariant** to nuisances
- Components of Z are as independent of each other as much as possible

Problem with disentanglement

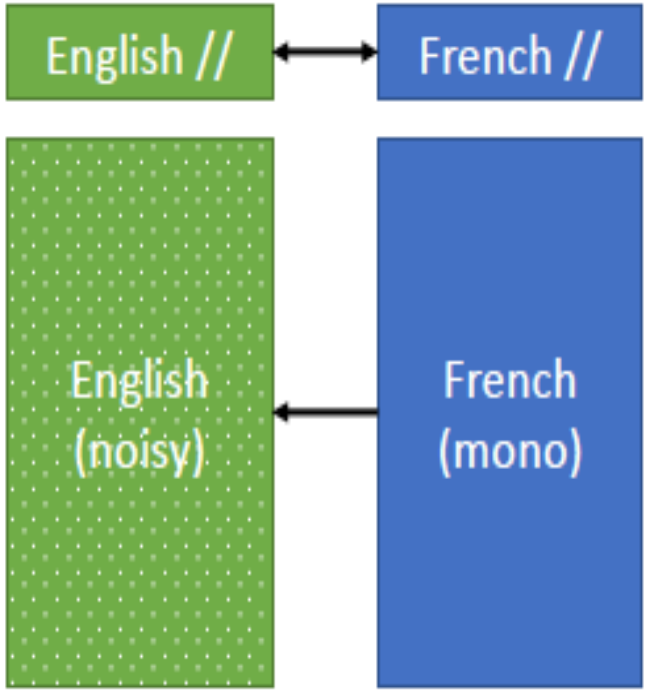
- We consider z to be disentangled from y if it is impossible to recover y from z.
- While failure to recover y from z could mean either that z was disentangled or that the classifier chosen to recover y was either not powerful enough or poorly trained,
- success of any classifier in recovering y demonstrates that z was in fact not invariant to y.

Approach

- Denoising auto-encoding - noise that corrupts the input sentence by performing word drops and word order shuffling. x_c is a corrupted version of the sentence x.
- Back translation - take an input (x,y) and encode x it into z, but then decode using another set of attribute values, $\tilde{y}$, yielding the reconstruction $\tilde{x}$. Use $\tilde{x}$ as input of the encoder and decode it using the original y to ideally obtain the original x, and train the model to map ($\tilde{x}$, y) into x

Back Translation

- Train system in the opposite direction
- Input sentences will be somewhat corrupt because of translation errors of the first system.
- Exposes the model to a pseudo-supervised setting, where the model's outputs act as supervised training data for the ultimate task at hand



Denoising Auto Encoder

- To reconstruct data from an input of corrupted data.
- After giving the autoencoder the corrupted data, the hidden layer is forced to learn only the more robust features, rather than just the identity.
- The output will then be a more refined version of the input data
- stochastically corrupt data sets and input them into a neural network.
- The autoencoder can then be trained against the original data.
- One way to corrupt the data would be simply to randomly remove some parts of the data, so that the autoencoder is trying to predict the missing input

Denoising Auto Encoder

- Train a source -> source denoising autoencoder (DAE)
- Add noise to avoid trivial reconstructions
  - Word drop out

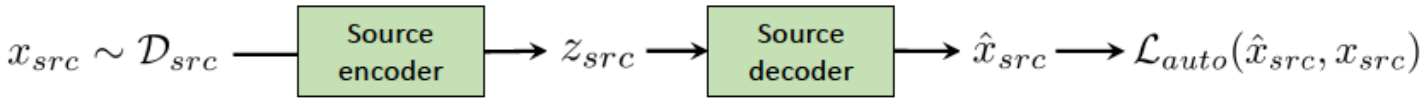
E. g. Arizona was the first to introduce such a requirement

Arizona was first to such a requirement

Arizona was first to introduce such a requirement

  - Word order shuffle

E. g.



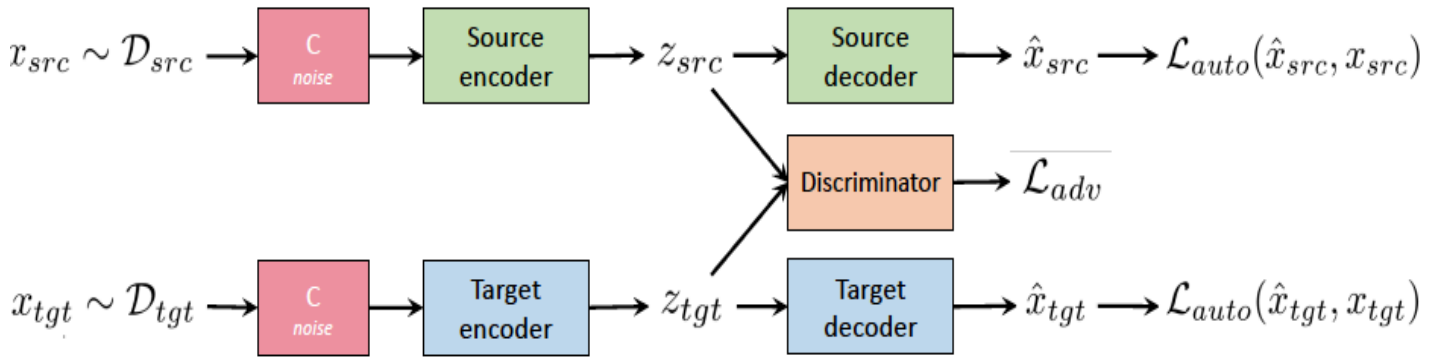
Arizona was the first to introduce such a requirement

Arizona was the first to introduce such a requirement

was Arizona first the was to such introduce requirement a

Denoising Auto Encoder Cont'd

- Train target -> target denoising autoencoder (DAE)
- Make source and target latent states indistinguishable using adversarial training
- Decoders should operate in same space -> share parameters between encoders



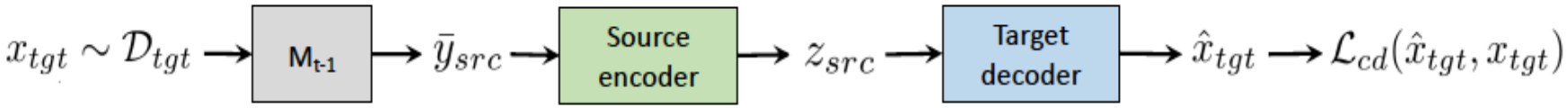
Cross Domain Training

- Train the system to do actual style transfer
  - No parallel data available – generate artificial data using back translation
- Train on pairs generated using a previous version of the model
  - Start with a word by word translation

X_{tgt} Sentence from monolingual corpus

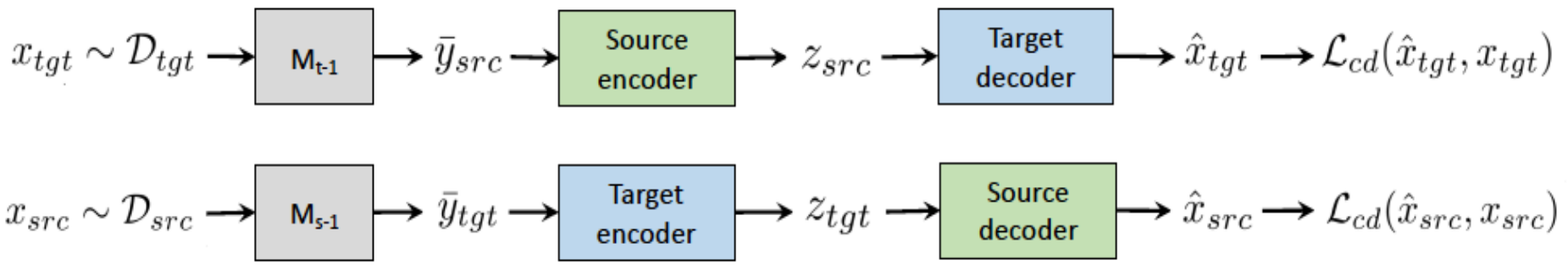
$\tilde{y}_{src}$ Word-by-word translation

Y_{src} Gold translation (by humans)



Cross Domain Training

- Train on pairs generated using a previous version of the model
- Train the model on generated sentence pairs like in a supervised setting



Recap

- Denoising autoencoder to learn good sentence representations
- Match distributions of latent features across the two domains
  - Adversarial training
  - Parameter sharing
- Cross-lingual training to learn to translate
  - Use current version of model to produce noisy source
- Pretrain word embeddings with aligned embeddings

Problems?

- Model swapping a single binary attribute
- Not enough control on the trade-off between content preservation and change of attributes.

Solution

- Attribute conditioning - separately embed each target attribute value and then average their embeddings. Then feed the averaged embeddings to the decoder as a start-of-sequence symbol.
- The output layer of the decoder uses a different bias for each attribute label
- Latent representation pooling to control the amount of content preservation.
  - The smaller the denoising term the more content is preserved and the less well attributes are swapped), the temperature T used to produce unbiased generations and to control the amount of content preservation, and the pooling window size w.

Datasets

- Granularity of entire reviews
- Relax constraints that discard reviews with more than 15 words and only consider the 10k most frequent words.
- Consider full reviews with up to 100 words, and byte-pair encodings (BPE) with 60k BPE codes, eliminating the presence of unknown words.
- Annotations for the gender of the review author and the category of the product or restaurant being reviewed.

Datasets

	Sentiment		Gender		Category				
SYelp	Positive 266,041	Negative 177,218	Male -	Female -	American -	Asian -	Bar -	Dessert -	Mexican -
FYelp	Positive 2,056,132	Negative 639,272	Male 1,218,068	Female 1,477,336	American 904,026	Asian 518,370	Bar 595,681	Dessert 431,225	Mexican 246,102
Amazon	Positive 64,251,073	Negative 10,944,310	- -	- -	Book 26,208,872	Clothing 14,192,554	Electronics 25,894,877	Movies 4,324,913	Music 4,574,167
Social Media Content	Relaxed 7,682,688	Annoyed 17,823,468	Male 14,501,958	Female 18,463,789	18-24 12,628,250	65+ 7,629,505			

Table 3: The number of reviews for each attribute for different datasets. The *SYelp*, *FYelp* and the Amazon datasets are composed of 443k, 2.7M and 75.2M sentences respectively. Public social media content is collected from 3 different data sources with 25.5M, 33.0M and 20.2M sentences for the *Feeling*, *Gender* and *Age* attributes respectively.

Datasets

- Yelp reviews
  - remove reviews not written in English (fastText); not about restaurants; rated 3/5 stars (neutral sentiment)
  - Labels: Asian, American, Mexican, Bars & Dessert,
- Amazon reviews
  - exception of collecting gender labels (username often absent)
  - Books, Clothing, Electronics, Movies, Music,
- Public social media
  - gender (male or female); age group (18-24 or 65+); writer-annotated feeling (relaxed or annoyed)

Detailed Architecture

- 2-layer bidirectional LSTM with attention
  - 512 hidden units
  - 512 hidden units for embedding attribute
  - Decoder conditions on attribute embeddings

EVALUATION

- Attribute control: measure the extent to which attributes are controlled using fastText classifiers, trained on our datasets, to predict different attributes.
- Fluency: Fluency is measured by the perplexity assigned to generated text sequences by a pre-trained Kneser–Ney smooth 5-gram language model using KenLM
- Content preservation: using n-gram statistics, by measuring the BLEU score between generated text and the input itself, which we refer to as self-BLEU.
- human evaluations public crowd-sourcing platform.

Results

Model	Accuracy	BLEU	PPL
Fader/StyleEmbedding (Fu et al., 2017)	18%	16.7	56.1
MultiDecoder (Fu et al., 2017)	52%	11.3	90.1
CAE (Shen et al., 2017)	72%	6.8	53.0
Retrieval (Li et al., 2018)	81%	1.3	7.4
Rule-based (Li et al., 2018)	73%	22.3	118.7
DeleteOnly (Li et al., 2018)	77%	14.5	67.1
DeleteAndRetrieve (Li et al., 2018)	79%	16.0	66.6
Fader (Ours w/o backtranslation & attention)	71%	15.7	35.1
Ours	87%	14.6	26.2
Ours	85%	24.2	26.5
Ours	74%	31.2	49.8
Input copy	13%	30.6	40.6

Table 4: Automatic evaluation of models on the *SYelp* test set from Li et al. (2018). The test set is composed of sentences that have been manually written by humans, which we use to compute the BLEU score. Samples for previous models were made available by Li et al. (2018).

	Fluency	Content	Sentiment
DAR (Li et al. (2018))	3.33 (1.39)	3.16 (1.43)	64.05%
Ours	4.07 (1.12)	3.67 (1.41)	69.66%
Human (Li et al. (2018))	4.56 (0.78)	4.01 (1.25)	81.35%
	Our Model	No Preference	DAR
DAR vs Our Fader	**37.6%**	32.7%	29.7%
DAR vs Ours	**54.4%**	24.7%	20.8%

Table 5: **Top:** Results from human evaluation to evaluate the fluency / content preservation and successful sentiment control on the Li et al. (2018) *SYelp* test set. The mean and standard deviation of Fluency and Content are measured on a likert scale from 1-5 while sentiment is measured by fraction of times that the controlled sentiment of model matches the judge's evaluation of the sentiment (when also presented with a neutral option). **Bottom:** Results from human A/B testing of different pairs of models. Each cell indicates the fraction of times that a judge preferred one of the models or neither of them on the overall task.)

Results

Dataset (Model)	Attributes	Sentiment		Category		Gender	
		Accuracy	self-BLEU	Accuracy	self-BLEU	Accuracy	self-BLEU
Yelp (Fader)	Sentiment	85.5%	31.3	-	-	-	-
	Sentiment + Category	85.1%	20.6	46.1%	22.6	-	-
	Sentiment + Category + Gender	86.6%	20.4	47.7%	22.5	58.5%	23.3
Yelp (Ours)	Sentiment	87.4%	54.5	-	-	-	-
	Sentiment + Category	87.1%	38.8	64.9%	44.0	-	-
	Sentiment + Category + Gender	88.5%	31.6	64.1%	36.5	59.0%	37.4
	Gender	-	-	-	-	59.1%	47.0
Amazon (Ours)	Sentiment	82.6%	54.8	-	-	-	-
	Sentiment + Category	82.5%	48.9	81.4%	41.8	-	-
Input Copy	-	50.0%	100.0	20.0%	100.0	50.0%	100.0

Table 6: Results using automatic evaluation metrics on the *FYelp* and *Amazon* test sets. Different rows correspond to the set of attributes being controlled by the model.

Model	Test (FYelp)		Test (Li et al., 2018)	
	Accuracy	self-BLEU	Accuracy	BLEU
Our model	87%	54.5	80%	25.8
-pooling	89%	47.9	-	-
-temperature	86%	45.2	80%	21.3
-attention	93%	25.4	80%	22.1
-back-translation	86%	32.8	69%	16.4
+adversarial	86%	45.5	78%	25.1
-attention -back-translation	90%	26.0	71%	15.7

Table 7: Model ablations on 5 model components on the *FYelp* dataset (Left) and *SYelp* (Right).

Results

Sentiment	Category	Input / Generations
		Amazon
Positive	**Movies**	**exciting new show. john malkovich is superb as always. great supporting cast. hope it survives beyond season 1**
Positive	Books	exciting new book. john grisham is one of the best. great read. hope he continues to write more.
Negative	Books	nothing new. john grisham is not as good as his first book. not a good read.
Positive	Clothing	awesome new watch. fits perfectly. great price. great quality. hope it lasts for a long time.
Negative	Clothing	horrible. the color is not as pictured. not what i expected. it is not a good quality.
Positive	Electronics	works great. the price is unbeatable. great price. great price. hope it lasts for a long time.
Negative	Electronics	worthless. the picture is not as clear as the picture. not sure why it is not compatible with the samsung galaxy s2.
Positive	Movies	exciting new show. john goodman is great as always. great supporting cast. hope it continues to end.
Negative	Movies	horrible. the acting is terrible. not worth the time. it's not worth the time.
Positive	Music	awesome new album. john mayer is one of the best. great album. hope he continues to release this album.
Negative	Music	horrible. the songs are not as good as the original. not worth the price.
		Yelp
Negative	**Dessert**	**the bread here is crummy, half baked and stale even when "fresh." i won't be back.**
Positive	American	the burgers here are juicy, juicy and full of flavor! i highly recommend this place.
Negative	American	the bread here is stale, dry and over cooked even though the bread is hard. i won't be back.
Positive	Asian	the sushi here is fresh, tasty and even better than the last. i highly recommend this place.
Negative	Asian	the noodles here are dry, dry and over cooked even though they are supposed to be "fresh." i won't be back.
Positive	Bar	the pizza here is delicious, thin crust and even better cheese (in my opinion). i highly recommend it.
Negative	Bar	the pizza here is bland, thin crust and even worse than the pizza, so i won't be back.
Positive	Dessert	the ice cream here is delicious, soft and fluffy with all the toppings you want. i highly recommend it.
Negative	Dessert	the bread here is stale, stale and old when you ask for a "fresh" sandwich. i won't be back.
Positive	Mexican	the tacos here are delicious, full of flavor and even better hot sauce. i highly recommend this place.
Negative	Mexican	the beans here are dry, dry and over cooked even though they are supposed to be "fresh." i won't be back.

Table 9: Demonstrations of our model's ability to control multiple attributes simultaneously on the *Amazon* dataset (top) and *FYelp* dataset (bottom). The first two columns indicate the combination of attributes that are being controlled, with the first row indicating a pre-specified input