Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer

Authors: Juncen Li, Robin Jia, He He, Percy Liang

Task

text attribute transfer

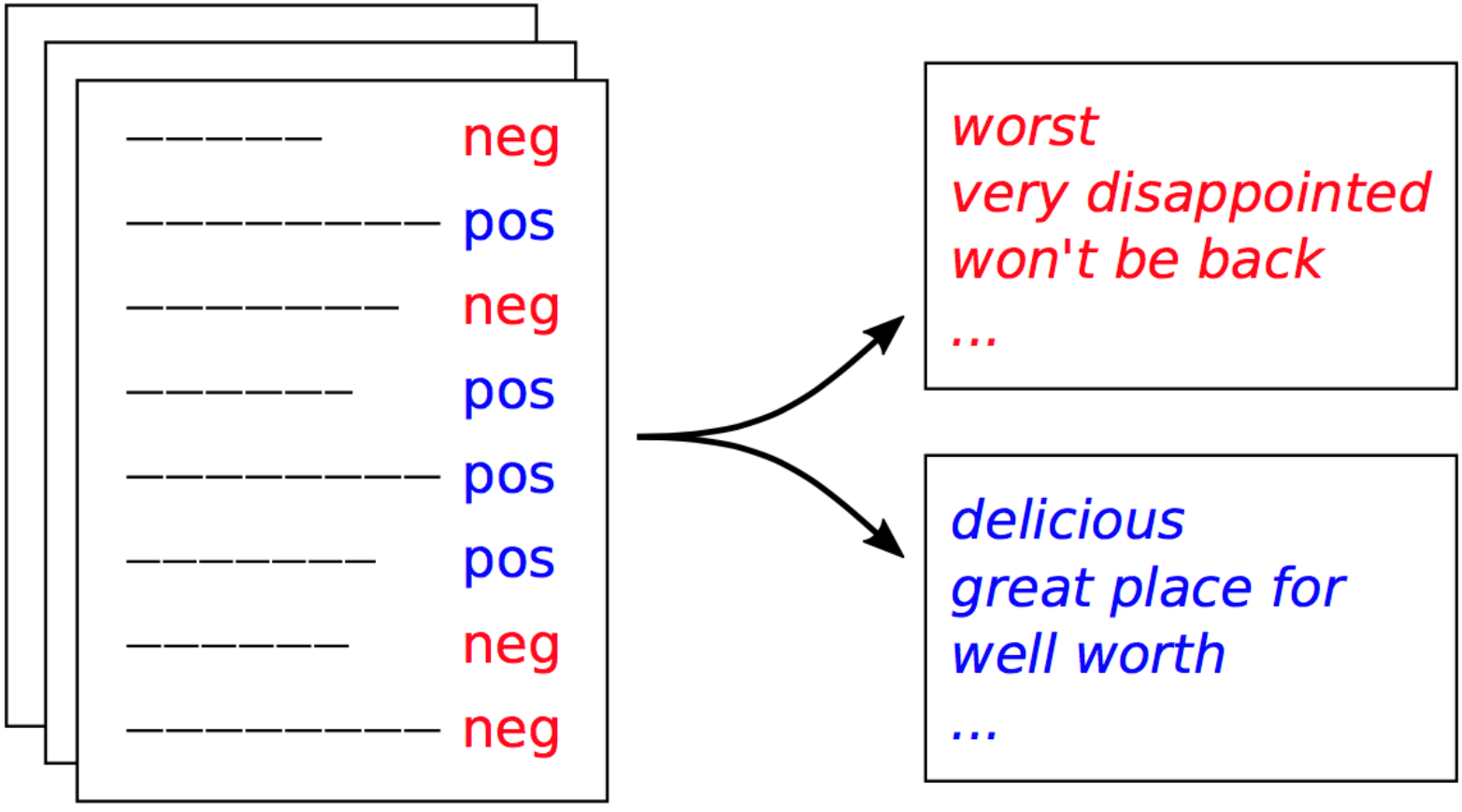
transforming a sentence to alter a specific attribute (e.g., sentiment) while preserving its attribute-independent content (e.g., changing “screen is just the right size” to “screen is too small”)

data

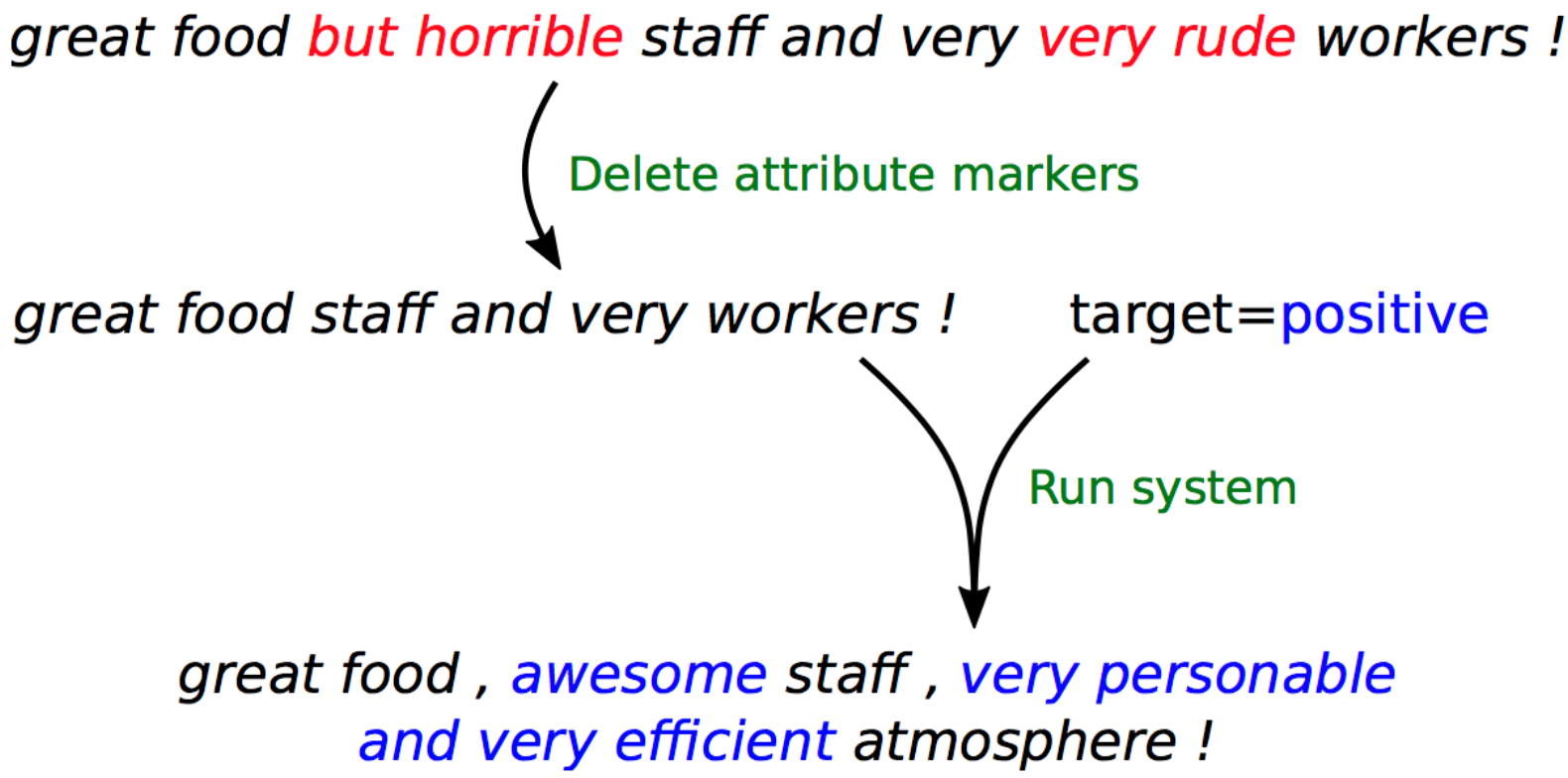
only sentences labeled with their attribute (e.g., positive or negative), but not pairs of sentences that differ only in their attributes

Method

attribute transfer can often be accomplished by changing a few attribute markers

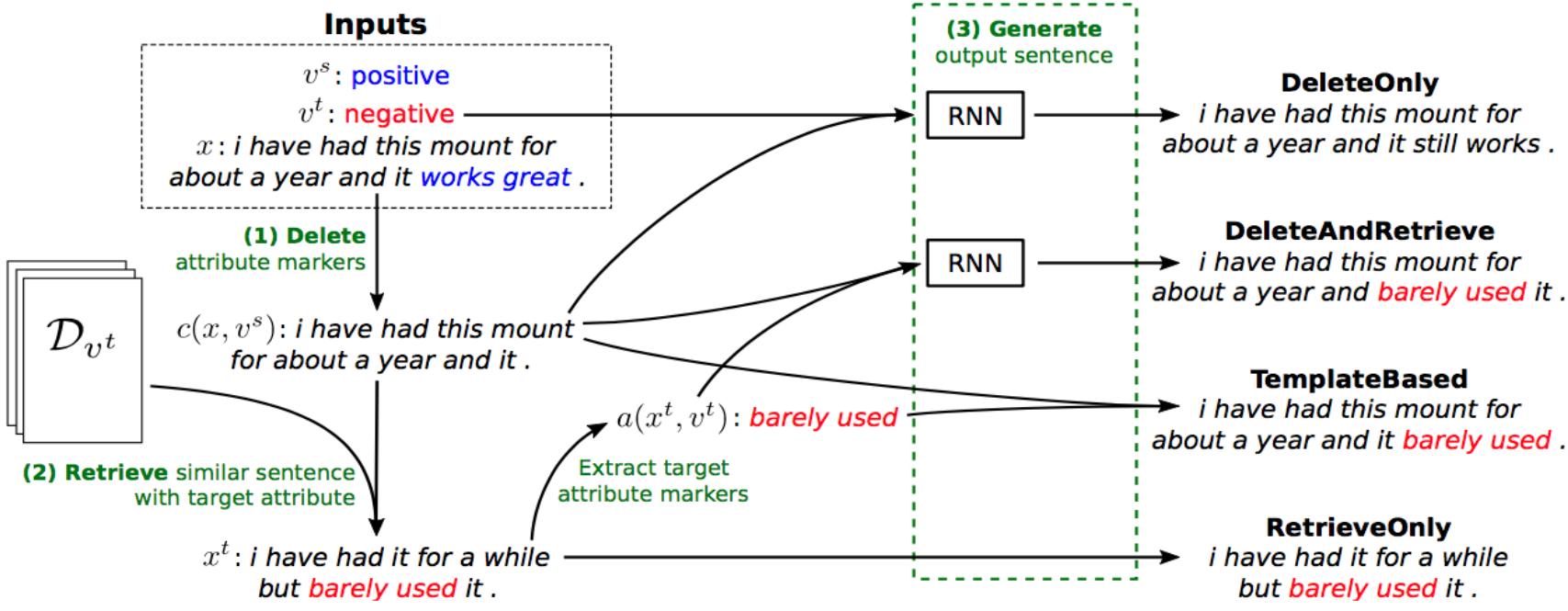


Method


great food but horrible staff and very very rude workers !
Delete attribute markers
great food staff and very workers !
target=positive
Run system
great food , awesome staff , very personable
and very efficient atmosphere !

Method

x : sentence / *v* : attribute / *c*: content / *a* : attribute maker



Method

- Delete attribute makers

  n-grams appear more often in source but less in target corpora

$$s(u, v) = \frac{\text{count}(u, \mathcal{D}_v) + \lambda}{\left(\sum_{v' \in \mathcal{V}, v' \neq v} \text{count}(u, \mathcal{D}_{v'})\right) + \lambda}$$

- Retrieve similar sentence with target attribute

  distance(TF-IDF / Euclidean distance) between contents in different attributes

- Generate

  DELETEONLY use an RNN content encoder with learned embeddings of attributes. *DELETEANDRETRIEVE* use another RNN to embed attribute makers of retrieved sentences instead

Dataset

1. YELP reviews
   positive or negative
2. AMAZON reviews
   positive or negative
3. changing image CAPTIONs to be romantic or humorous
   romantic, humorous & factual

Dataset	Attributes	Train	Dev	Test
YELP	Positive	270K	2000	500
	Negative	180K	2000	500
CAPTIONS	Romantic	6000	300	0
	Humorous	6000	300	0
	Factual	0	0	300
AMAZON	Positive	277K	985	500
	Negative	278K	1015	500

Human Reference

- Task: *flip sentences' its sentiment while preserving content*
- Goal: *understand the extent to which humans follow attribute maker pattern*
- indicator

1. *words marks as content & preserved by humans*

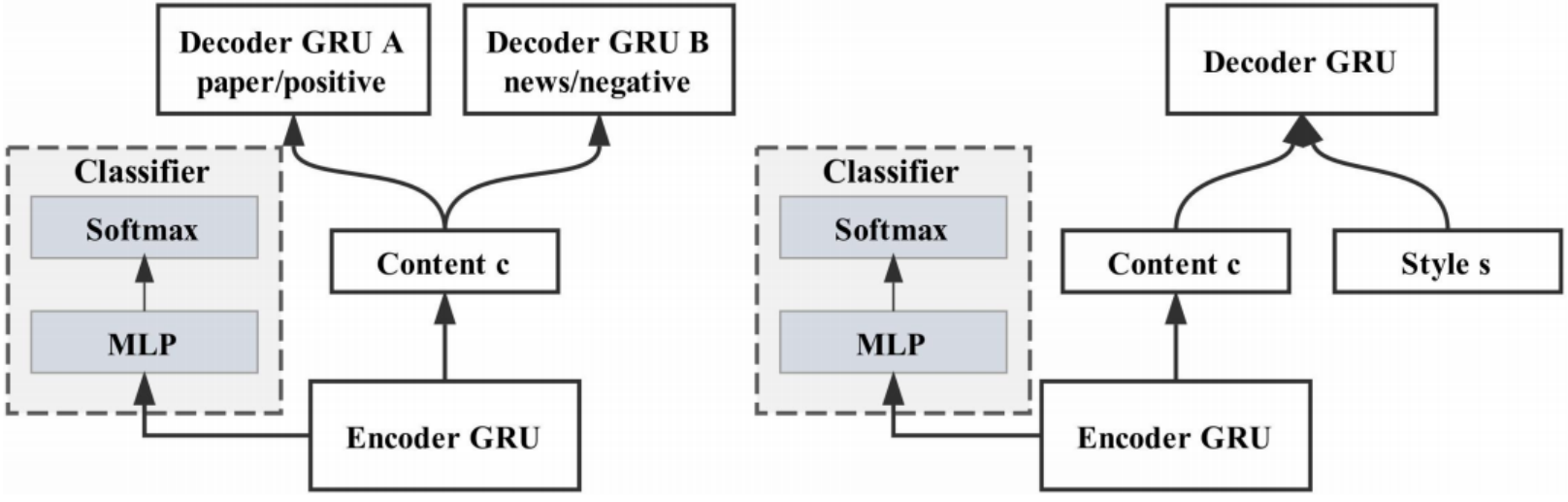
$$S_c = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x, v^{\text{src}}, y^*) \in \mathcal{D}_{\text{test}}} \frac{|c(x, v^{\text{src}}) \cap y^*|}{|c(x, v^{\text{src}})|}$$

2. *words marks as attribute-maker & changed by humans*

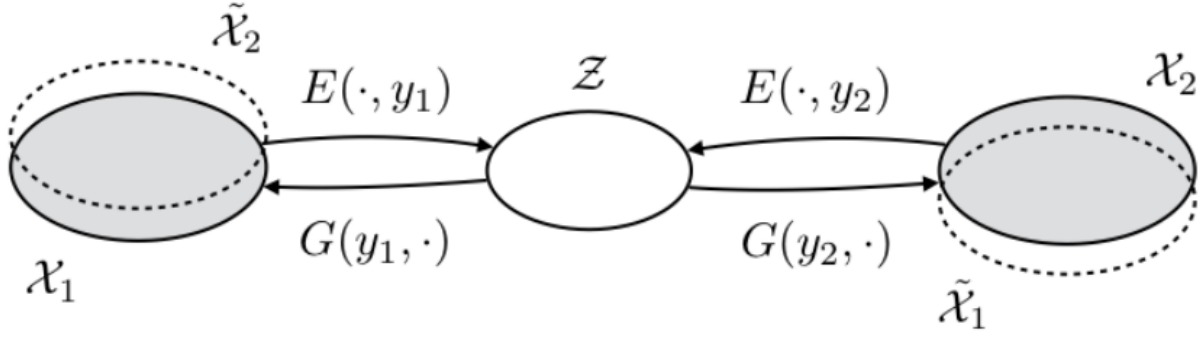
$$S_a = 1 - \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x, v^{\text{src}}, y^*) \in \mathcal{D}_{\text{test}}} \frac{|a(x, v^{\text{src}}) \cap y^*|}{|a(x, v^{\text{src}})|}$$

Baselines

1. STYLEEMBEDDING & MULTIDECODER



2. CROSSALIGNED



Human Evaluation

- 5 points Likert scale
- Grammaticality (Gra)
- Content preservation (Con)
- Target attribute match (Att)
- Success (Suc): rated 4 or 5 on all three criteria

	Yelp				Amazon				Captions			
	Gra	Con	Att	Suc	Gra	Con	Att	Suc	Gra	Con	Att	Suc
CrossAligned	2.8	2.9	3.5	14%	3.2	2.5	2.9	7%	3.9	2.0	3.2	16%
StyleEmbedding	3.5	3.7	2.1	9%	3.2	2.9	2.8	11%	3.3	2.9	3.0	17%
MultiDecoder	2.8	3.1	3.0	8%	3.0	2.6	2.8	7%	3.4	2.8	3.2	18%
RetrieveOnly	**4.2**	2.7	**4.2**	25%	3.8	2.8	3.1	17%	**4.2**	2.6	3.8	27%
TemplateBased	3.0	**3.9**	3.9	21%	3.4	3.6	3.1	19%	3.3	**4.1**	3.5	33%
DeleteOnly	3.0	3.7	3.9	24%	3.7	**3.8**	3.2	24%	3.6	3.5	3.5	32%
DeleteAndRetrieve	3.3	3.7	4.0	**29%**	**3.9**	3.7	**3.4**	**29%**	3.8	3.5	**3.9**	**43%**
Human	4.6	4.5	4.5	75%	4.2	4.0	3.7	44%	4.3	3.9	4.0	56%

Automatic Evaluation

- Target attribute match

  Attribute classifier trained on same data

- Content preservation

  BLEU scores with Human Reference

	Yelp		Captions		Amazon	
	Classifier	BLEU	Classifier	BLEU	Classifier	BLEU
CrossAligned	73.7%	3.1	74.3%	0.1	**74.1%**	0.4
StyleEmbedding	8.7%	**11.8**	54.7%	6.7	43.3%	10.0
MultiDecoder	47.6%	7.1	68.5%	4.6	68.3%	5.0
TemplateBased	81.7%	**11.8**	92.5%	**17.1**	68.7%	**27.1**
RetrieveOnly	**95.4%**	0.4	**95.5%**	0.7	**70.3%**	0.9
DeleteOnly	85.7%	7.5	83.0%	9.0	45.6%	24.6
DeleteAndRetrieve	88.7%	8.4	**96.8%**	7.3	48.0%	22.8

Corelation between Evaluations

	Classifier	BLEU	
	Attribute	Content	Grammaticality
All data	$0.810\ (p < 0.01)$	$0.876\ (p < 0.01)$	$-0.127\ (p = 0.58)$
YELP	$0.991\ (p < 0.01)$	$0.935\ (p < 0.01)$	$0.119\ (p = 0.80)$
CAPTIONS	$0.982\ (p < 0.01)$	$0.991\ (p < 0.01)$	$-0.631\ (p = 0.13)$
AMAZON	$-0.036\ (p = 0.94)$	$0.857\ (p < 0.01)$	$0.306\ (p = 0.50)$

Thank you!

Q&A