Text Style Transfer with Confounders

Tianxiao Shen Regina Barzilay Tommi Jaakkola
tianxiao@mit.edu

MIT CSAIL

What Is “Style Transfer”?

source target

Monet ⟶ photo

horse ⟶ zebra

[Zhu et al. 2017]

From informal to formal

Gotta see both sides of the story
→ You have to consider both sides of the story

[Rao et al. 2018]

From Shakespeare to modern

Send thy man away → Send your man away

[Xu et al. 2012]

From negative to positive sentiment

I would recommend find another place.
→ I would recommend this place again!

[Shen et al. 2017]

From dialect to written standard

From complex to simple sentences

...

MIT CSAIL

Easy: Paired Training Sets

- Supervised learning using paired examples of style transfer

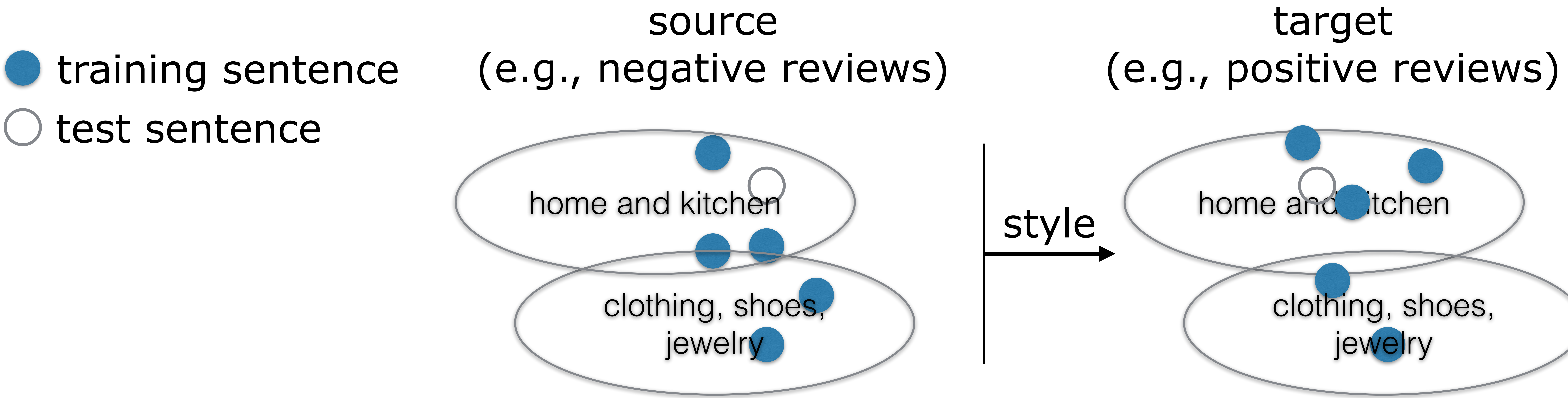
source (e.g., negative reviews)		target (e.g., positive reviews)
owner: a very rude man.	→	owner: a very friendly man.
i would not recommend giving them a try!	→	i'd definitely recommend giving them a try!
we were both so disappointed!	→	we were both so impressed!
consistently slow.	→	consistently fast.

- To collect parallel data is very costly or even impossible

MIT CSAIL

Intermediate: Unpaired but Distributionally Matched Sets

- Available source and target sentences as sets differ only in terms of style, i.e., they are distributionally matched otherwise



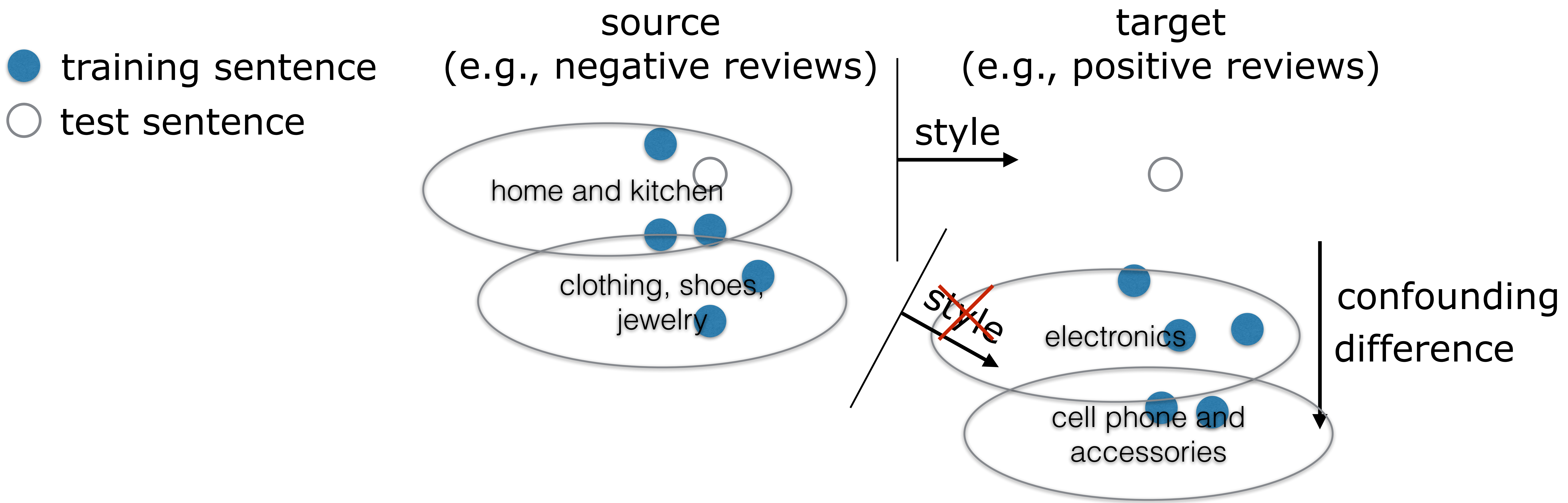
- The desired style change is just the source vs target difference
- New sentences map to sentences similar to those already seen during training

MIT CSAIL

[Shen et al. 2017]

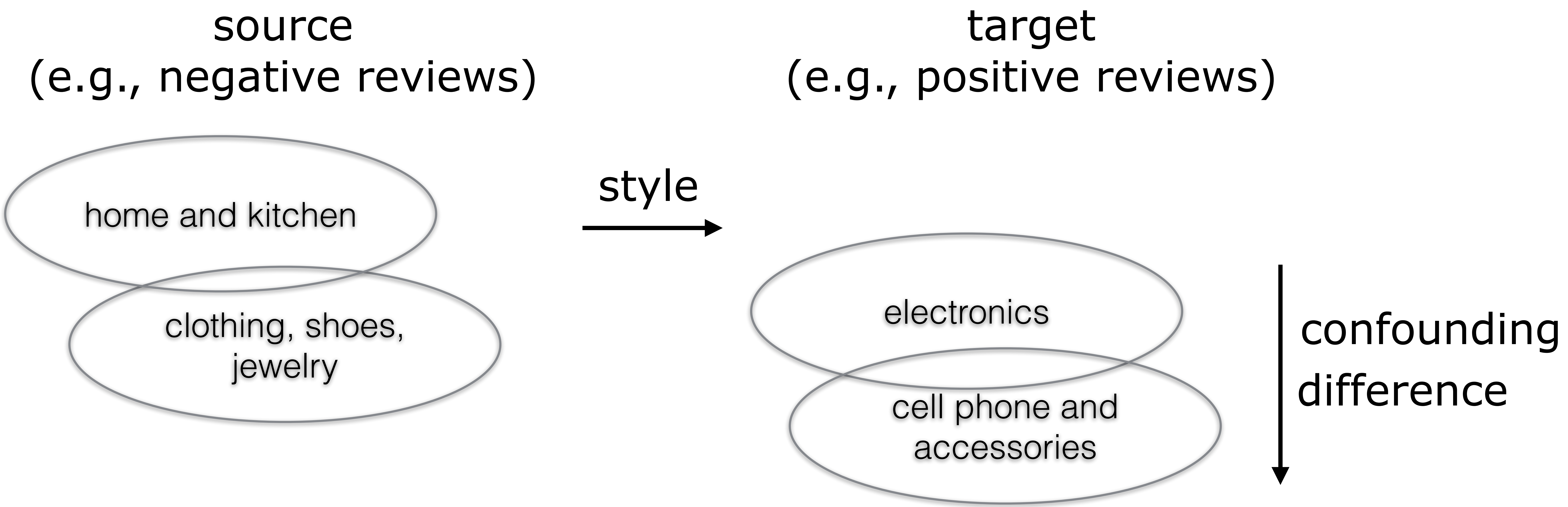
Hard: Unpaired, Not Distributionally Matched Sets

- There are **additional confounding differences** between source and target sentences



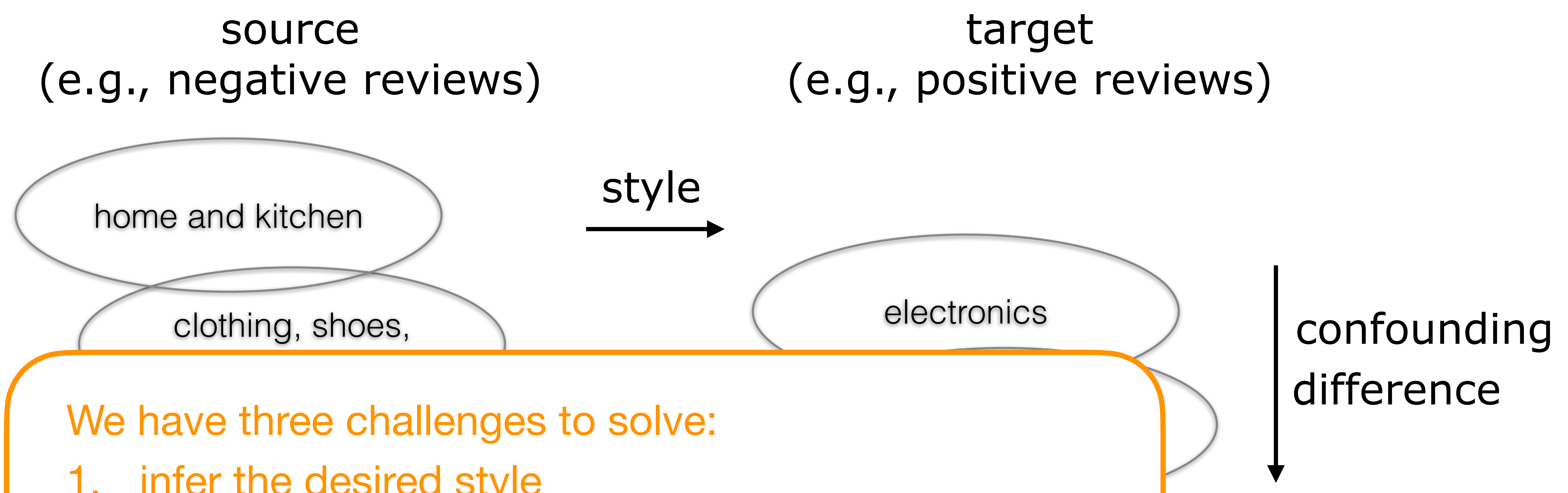
- Style change no longer equals source vs target difference
- New sentences map to type of sentences not seen during training

Solving Style Transfer with Confounders



- The task is illustrated by two groups of datasets (negative group and positive group), the **primary distinction** between them (sentiment) specifies the style to be transferred
- The intra-group variations (category) are **confounding differences** which need to be differentiated from the style and preserved during transfer

Solving Style Transfer with Confounders



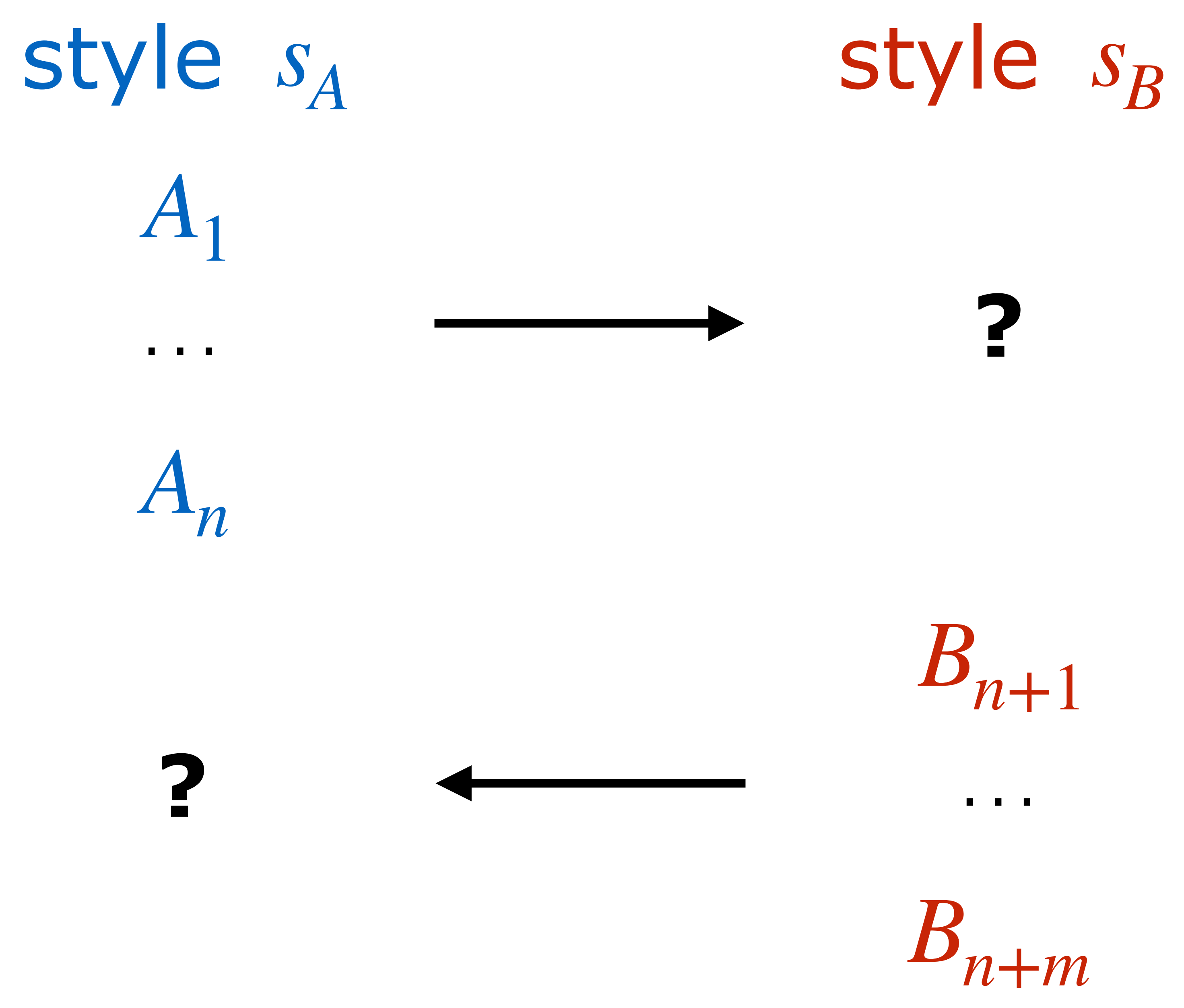
We have three challenges to solve:

1. infer the desired style
2. preserve aspects other than the inferred style
3. generate type of sentences not seen during training

- The tas… positive group), the **primar…** be transferred
- The intra-group variations (category) are **confounding differences** which need to be differentiated from the style and preserved during transfer

Task Formulation

- Given $A_1, \ldots, A_n$ of style s_A and $B_{n+1}, \ldots, B_{n+m}$ of style s_B, where A_i / B_j is a corpus consisting of sentences x
- Each corpus has its own characteristics
- Change only style and keep other aspects intact



MIT CSAIL

Model Overview

1. Learn a pair of classifiers to detect style and orthogonal attributes
   - Build on invariant risk minimization
2. Use the classifiers to guide a model to transfer in the desired direction

MIT CSAIL

1.0 Invariant Risk Minimization (IRM)

- Specify a set of environments $\mathscr{E} = \{e_1, \ldots, e_K\}$, where $e_k = \{(x_k^{(i)}, y_k^{(i)})\}_{i=1}^{n_k}$
- Environment difference accounts for nuisance variation we should **not** pay attention to
- Learn a feature representation that enables the same classifier to be optimal for all environments

- IRMv1: minimize empirical loss across all the data while penalize per-environment gradients with respect to any multiplier of the classifier output

$$\min_{\Phi:\mathscr{X}\to\mathscr{Y}} \sum_{e\in\mathscr{E}} R^e(\Phi) + \lambda \cdot \|\nabla_{w|w=1.0} R^e(w\cdot\Phi)\|^2$$

$R^e(\Phi) := \mathbb{E}_{X^e,Y^e}[\ell(\Phi(X^e), Y^e)]$

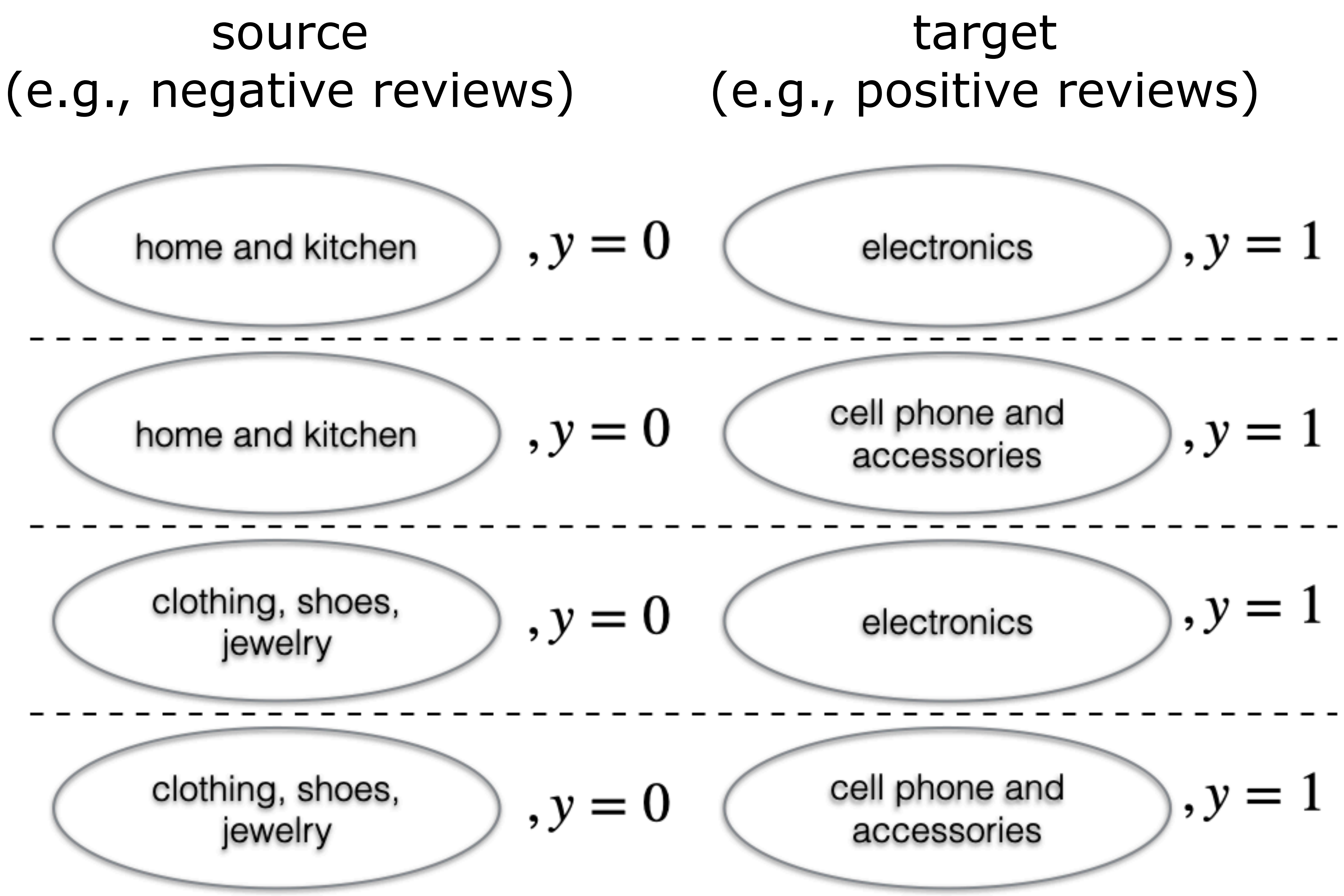
is the risk under environment e

gradients would be zero if Φ is per-environment optimal

MIT CSAIL

1.1 Inferring Style

- Construct environments $e_{i,j} = \{(x, y = 0) \mid x \in A_i\} \cup \{(x, y = 1) \mid x \in B_j\}$
- Learn IRM classifier $C_s : \mathcal{X} \to \mathcal{Y}$ across $\{e_{1,n+1}, \ldots, e_{n,n+m}\}$

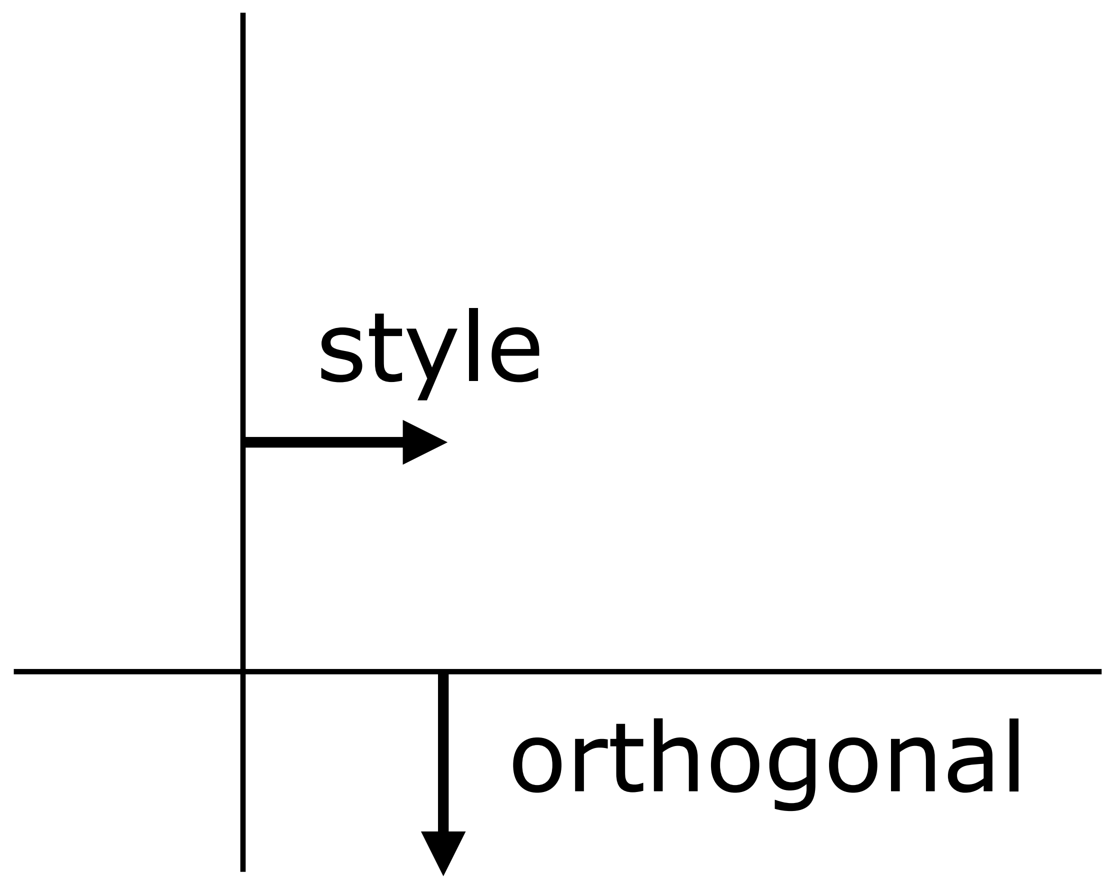


- Since all A_i share style s_A and all B_j share style s_B, style feature elicits an invariant classifier across $\{e_{i,j}\}$
- Conversely, if C_s uses any features specific to A_i/B_j, it won't be optimal in other $e_{i',j'}$

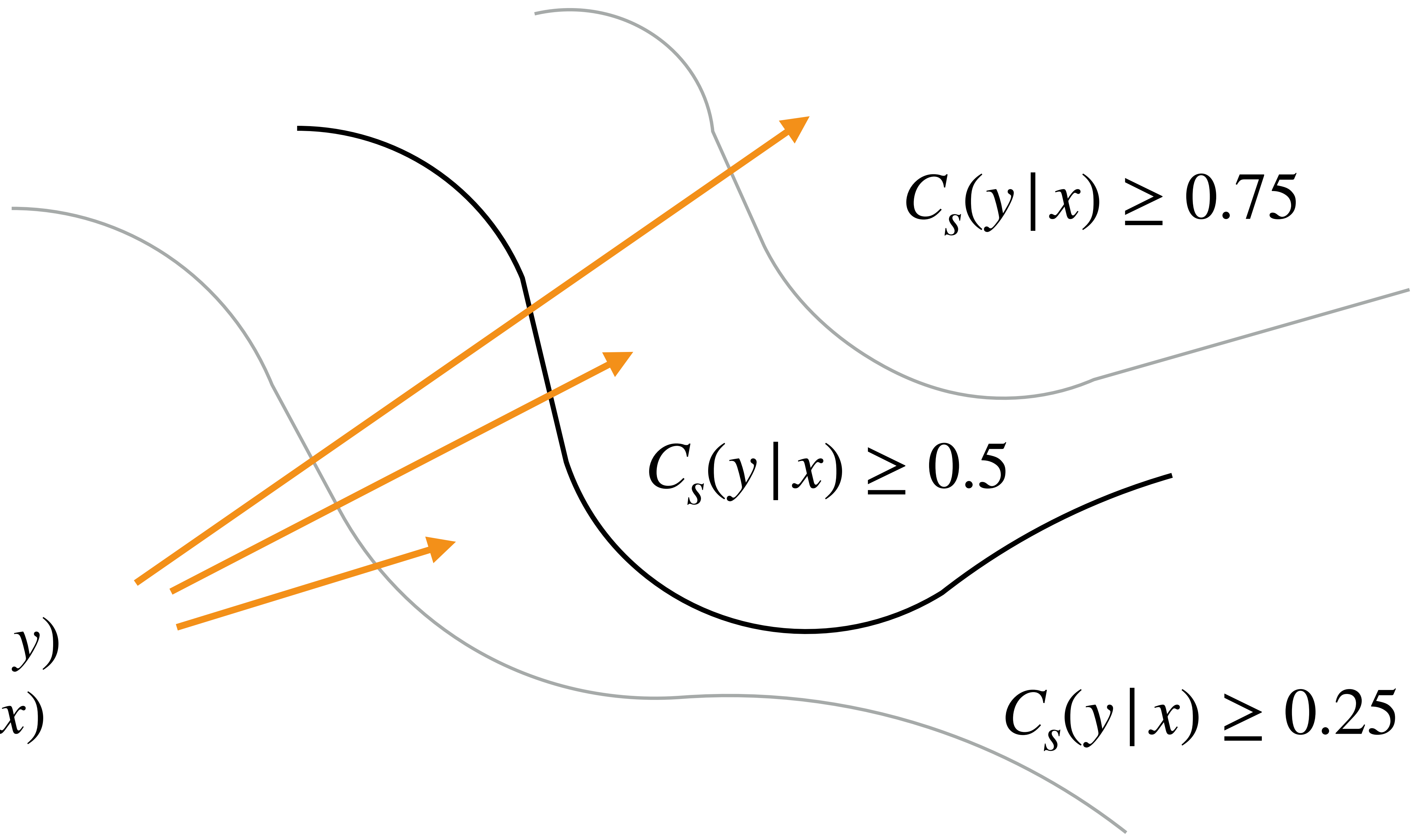
1.2 Inferring Style-Independent Aspects

- Let $A = A_1 \cup \ldots \cup A_n$, $B = B_{n+1} \cup \ldots \cup B_{n+m}$, $D = \{(x, y = 0) \mid x \in A\} \cup \{(x, y = 1) \mid x \in B\}$
- Construct environments based on C_s:

  $e_1 = \{(x, y) \in D \mid C_s(y \mid x) > 0.5\}$, $e_2 = \{(x, y) \in D \mid C_s(y \mid x) \leq 0.5\}$
- Learn IRM classifier $C_o : \mathcal{X} \to \mathcal{Y}$ across $\{e_1, e_2\}$



$C_o(y \mid x)$ is invariant
across contours of (x, y)
with respect to $C_s(y \mid x)$



2. Algorithm for Style Transfer

- Learn $M : \mathscr{X} \times \mathscr{Y} \rightarrow \mathscr{X}$ that takes a source sentence x and a target group y as input, and outputs a revised sentence that conforms to the style of group y
- Given a data example (x, y), let $\underline{\tilde{x} \sim M(x, 1-y)}$ be the transferred output
  - $\mathscr{L}_{rec} = -\log p_M(x \mid x, y)$ (reconstruction)
  - $\mathscr{L}_{C_s} = -\log p_{C_s}(1-y \mid \tilde{x})$ (different style)
  - $\mathscr{L}_{C_o} = -\log p_{C_o}(y \mid \tilde{x})$ (same orthogonal attributes)
  - $\mathscr{L}_{LM} = \underline{D_{KL}(p_M(\cdot \mid x, 1-y) \| p_{LM})}$ (language model regularization)
  - $\mathscr{L}_{BT} = -\log p_M(x \mid \tilde{x}, y)$ maximize entropy (back-translation)

use Gumbel-Softmax to back-propagate
- temperature annealing
- length control

$$\mathbb{E}_{(x,y)}[\mathscr{L}_{rec} + \lambda_1 \mathscr{L}_{C_s} + \lambda_2 \mathscr{L}_{C_o} + \lambda_3 \mathscr{L}_{LM} + \lambda_4 \mathscr{L}_{BT}]$$

MIT CSAIL

Baselines

- M with C_s: without C_o

$$\mathbb{E}_{(x,y)}[\mathscr{L}_{rec} + \lambda_1 \mathscr{L}_{C_s} \cancel{+ \lambda_2 \mathscr{L}_{C_o}} + \lambda_3 \mathscr{L}_{LM} + \lambda_4 \mathscr{L}_{BT}]$$

MIT CSAIL

Baselines

- M with C_s: without C_o
- M with C_{ERM}: guided by ERM classifier between A and B instead of C_s and C_o

$$+\lambda\mathscr{L}_{C_{ERM}} = -\log p_{C_{ERM}}(1-y\,|\,\tilde{x})$$

$$\boxed{\mathbb{E}_{(x,y)}[\mathscr{L}_{rec} \cancel{+\lambda_1\mathscr{L}_{C_s} + \lambda_2\mathscr{L}_{C_o}} + \lambda_3\mathscr{L}_{LM} + \lambda_4\mathscr{L}_{BT}]}$$

Baselines

- M with C_s: without C_o
- M with C_{ERM}: guided by ERM classifier between A and B instead of C_s and C_o
- He et al. (2020): regard non-parallel data as partially observed parallel data; treat transferred sequences as latent variables and derive ELBO

$$D_{KL}(p_M(\cdot \mid x, 1-y) \| p_{LM_{1-y}})$$

$$\boxed{\mathbb{E}_{(x,y)}[\mathscr{L}_{rec} \cancel{+ \lambda_1 \mathscr{L}_{C_s} + \lambda_2 \mathscr{L}_{C_o}} + \lambda_3 \mathscr{L}_{LM} + \lambda_4 \mathscr{L}_{BT}]}$$

MIT CSAIL

Baselines

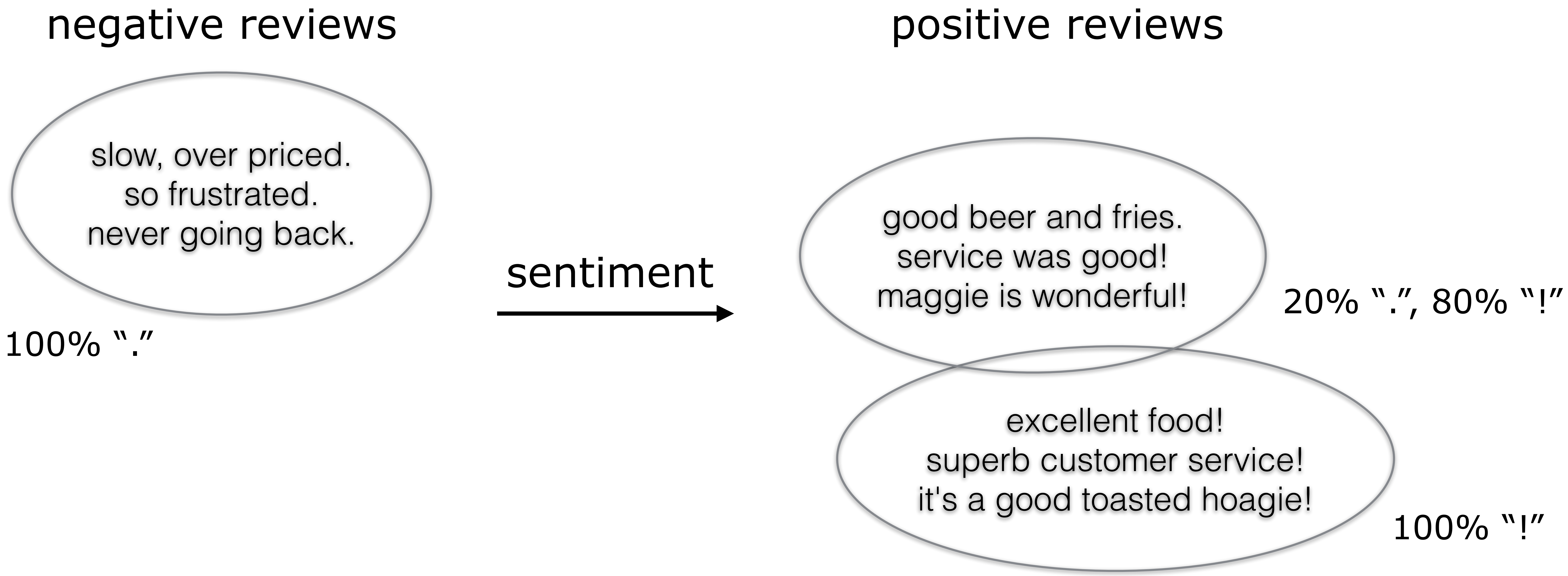
- M with C_s: without C_o
- M with C_{ERM}: guided by ERM classifier between A and B instead of C_s and C_o
- He et al. (2020): regard non-parallel data as partially observed parallel data; treat transferred sequences as latent variables and derive ELBO
- Krishna et al. (2020): use a separate paraphrasing dataset D_{pp}
    1. train M on D_{pp}, and use it to paraphrase A to A', B to B'
    2. train inverse models M_A to map A' to A, M_B to map B' to B
    3. to transfer a sentence to style A/B, apply M and then M_A/M_B

- D_{pp} needs to exclude unwanted changes
- D_{pp} needs to cover the desired style transformation, otherwise the models are applied OOD

MIT CSAIL

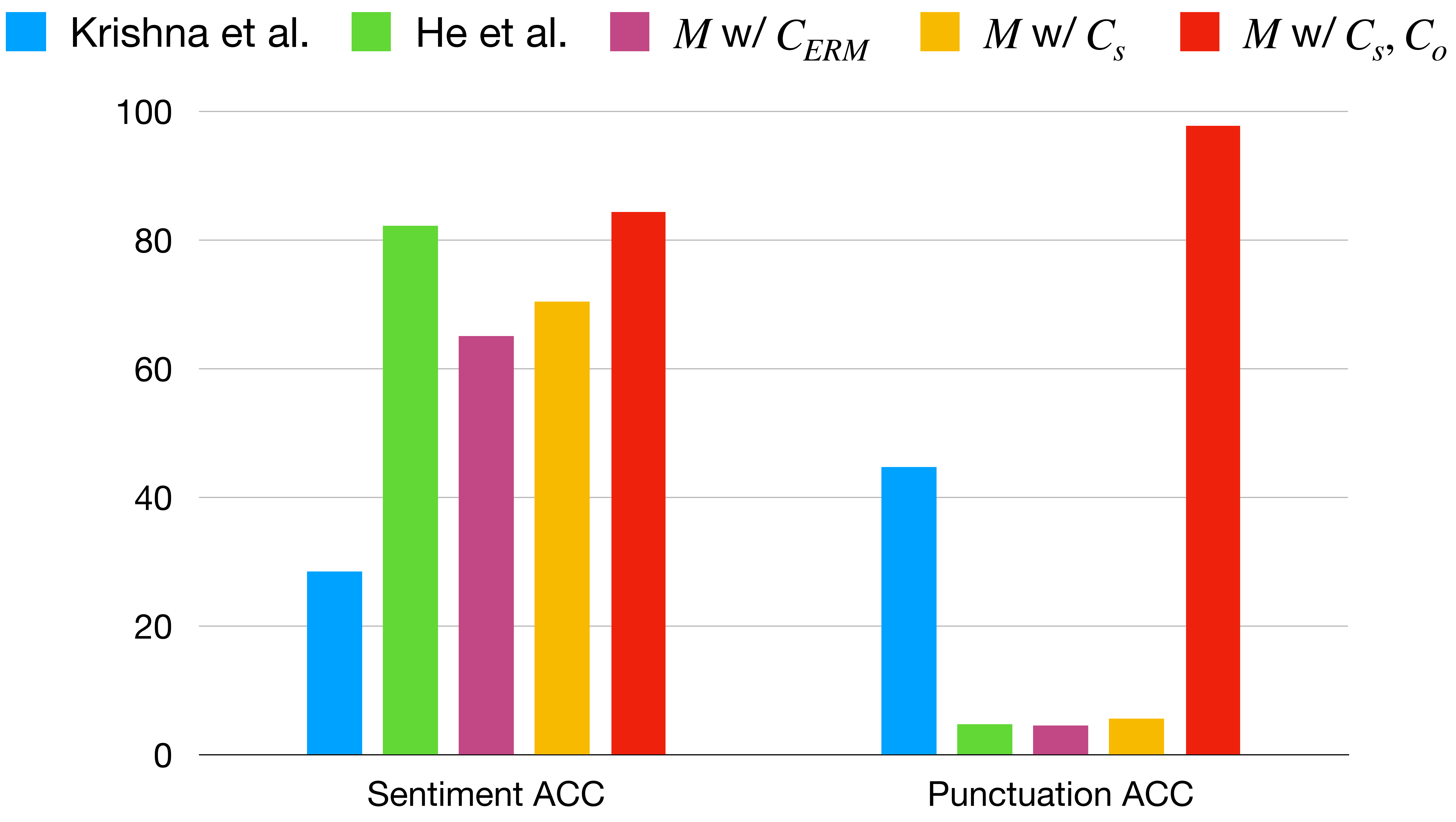
Sentiment Transfer with Different Punctuations

- Adapt sentiment transfer dataset, modifying punctuation to create spurious correlation
- Goal: alter sentiment without changing punctuation



MIT CSAIL

Automatic Evaluation Results



MIT CSAIL

Example Outputs

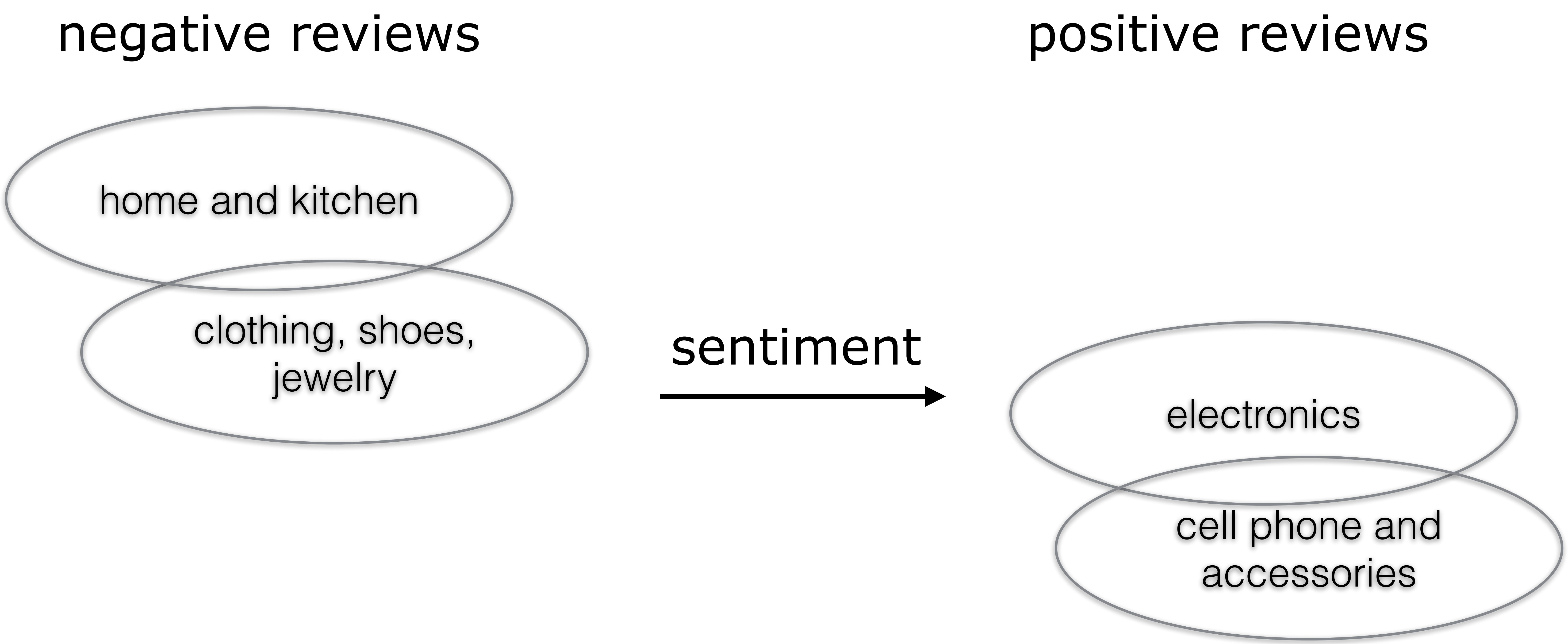
Input	the sales people here are terrible .
Krishna et al.	the people here are absolutely terrible .
He et al.	the sales people here are great !
M w/ C_{ERM}	the sales people here are amazing !
M w/ C_s	the sales people here are fantastic !
M w/ C_s, C_o (Ours)	the sales people here are amazing .

Input	excellent combination of flavors , very unique !
Krishna et al.	very unique combination of flavors , very unique ! ” .
He et al.	horrible customer service .
M w/ C_{ERM}	terrible combination of flavors , very disappointing .
M w/ C_s	terrible combination of flavors , not unique .
M w/ C_s, C_o (Ours)	terrible combination of flavors , not outstanding !

MIT CSAIL

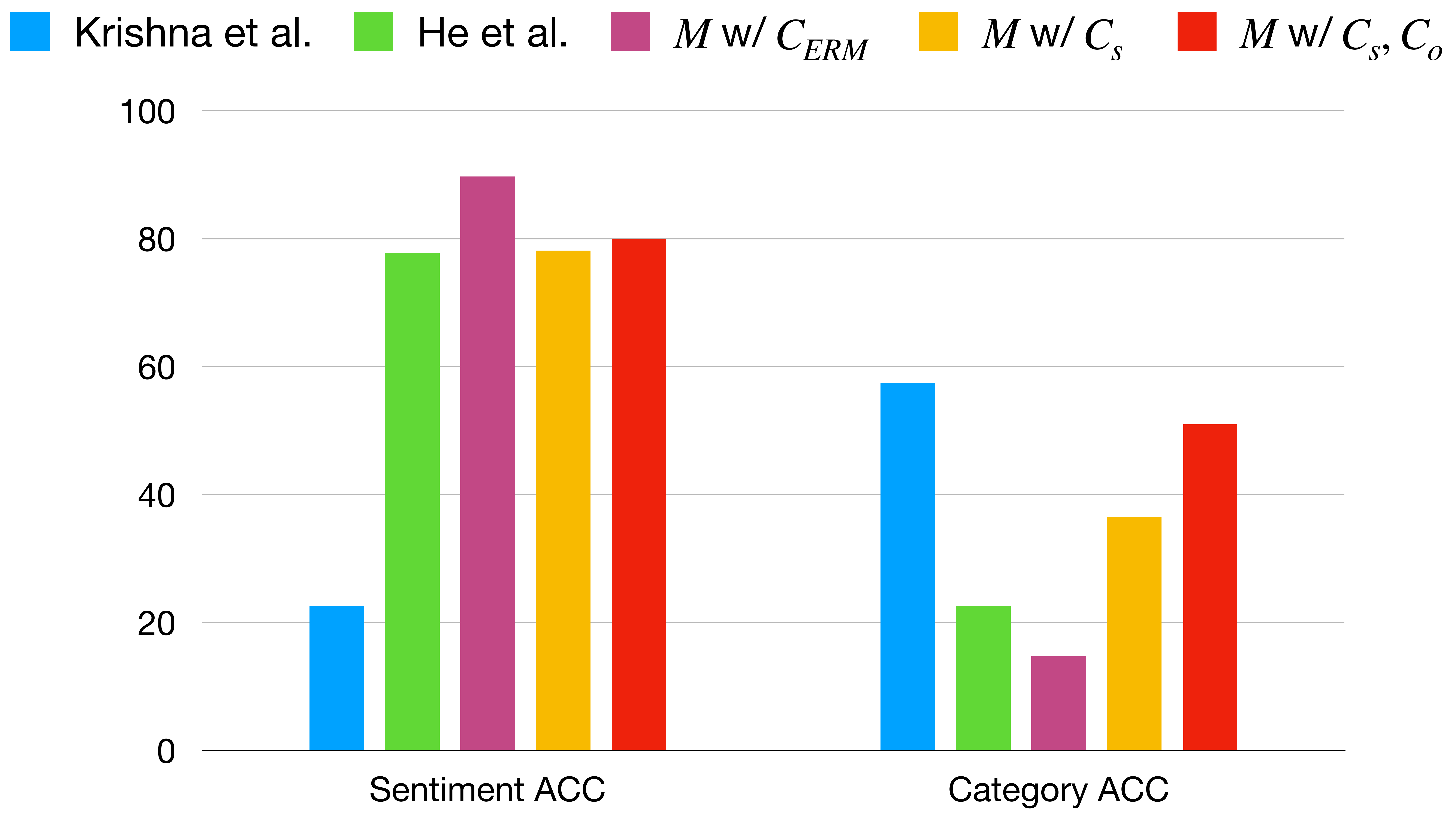
Sentiment Transfer with Different Categories

- Take positive and negative Amazon reviews from different categories
- Goal: alter sentiment without changing product category

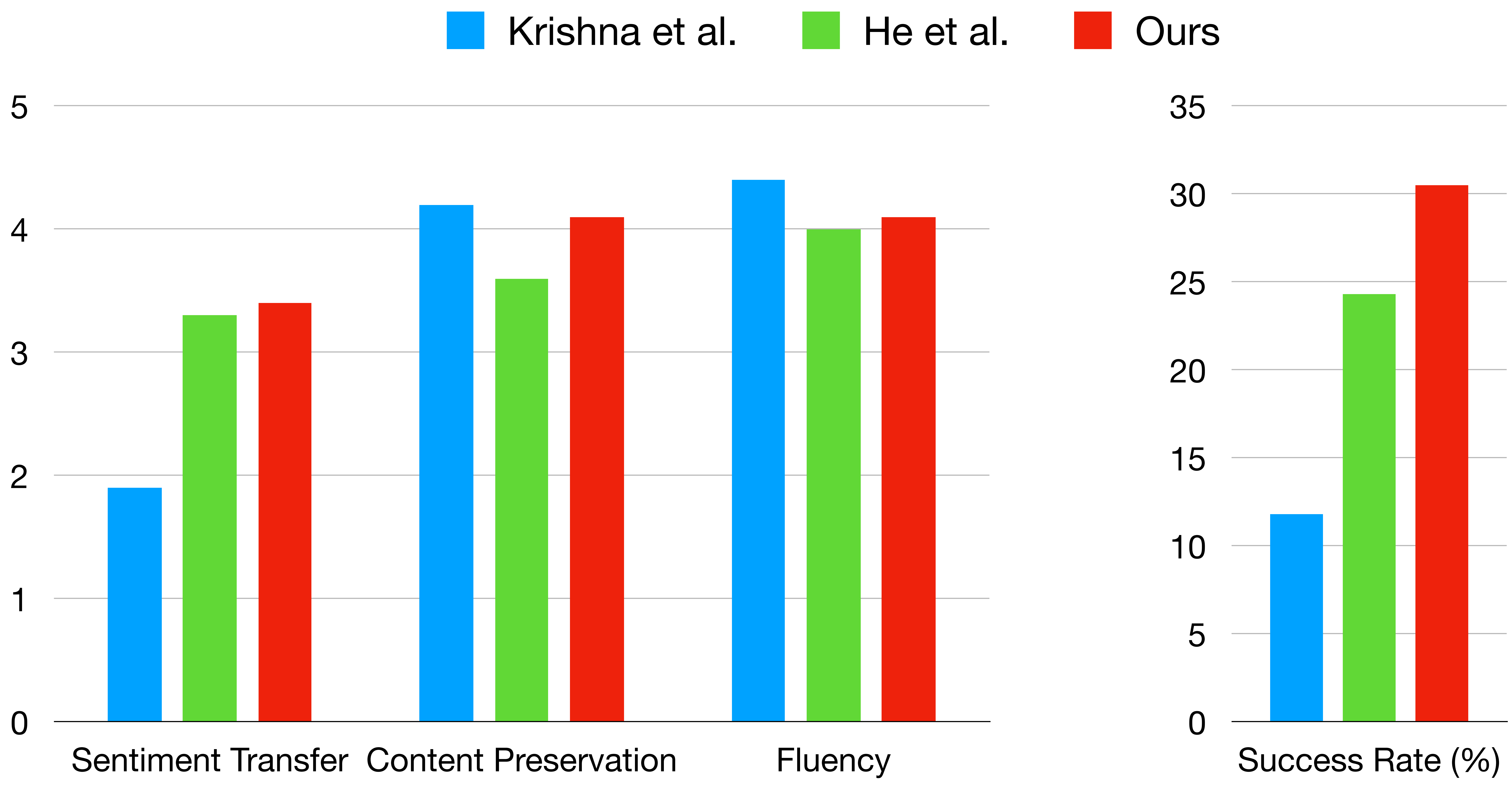


MIT CSAIL

Automatic Evaluation Results



Human Evaluation Results



Example Outputs

Input	this shirt was too tight . the sizing seems off .
Krishna et al.	the shirt is too tight .
He et al.	this case was great . the protection seems great .
Ours	this shirt works just perfect . the sizing seems well .
Input	the containers do not lock well and are made of low quality materials .
Krishna et al.	the containers do not fit securely and are made from poor quality material .
He et al.	the phones work well and has made of sound quality of low quality materials .
Ours	the containers does the job well and are made of high quality materials .
Input	exactly as advertised . converted a molex plug into a sata
Krishna et al.	the molex plug was convert to sata as advertised .
He et al.	way too big . leaves a inaccurate cut into a bath
Ours	not as advertised . converted a molex plug into a sata

MIT CSAIL

A Step Forward: An Aspirational Example

- Transfer from sonnets to tweets (author is a confounder)

source

target

Shakespeare's sonnets
Browning's sonnets
Pushkin's sonnets
…

→ ?

(Shakespeare's tweets)

Obama's tweets
Bieber's tweets
Perry's tweets
…

MIT CSAIL