SwitchOut: An Efficient Data Augmentation for Neural Machine Translation

Xinyi Wang*, Hieu Pham*, Zihang Dai, Graham Neubig

Carnegie Mellon University
Language Technologies Institute

November 2, 2018

*:equal contribution

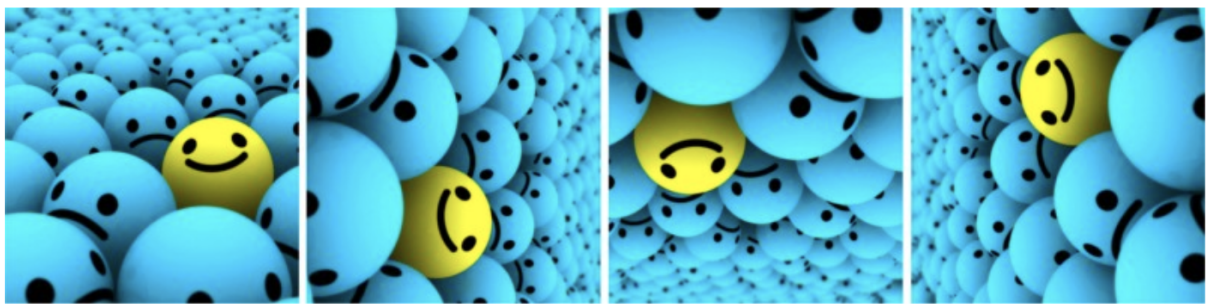
Data Augmentation

- Neural models are data hungry, while collecting data is expensive

[1] image source:Medium

Data Augmentation

- Neural models are data hungry, while collecting data is expensive
- Prevalent in computer vision[1]

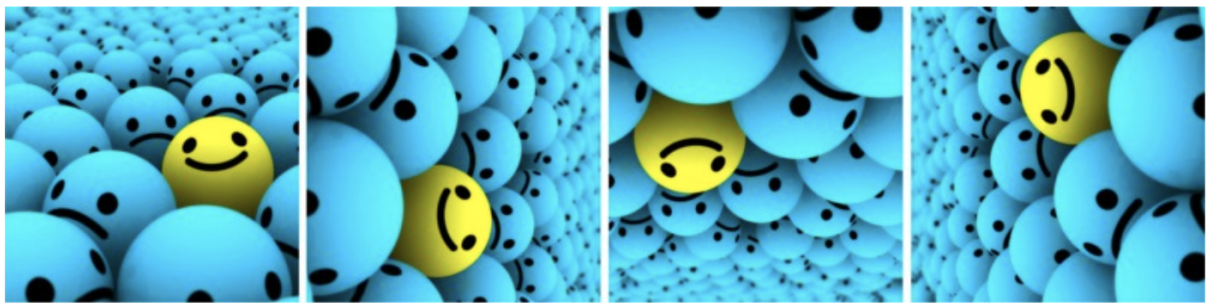


[1] image source:Medium

Data Augmentation

Language Technologies Institute

- Neural models are data hungry, while collecting data is expensive
- Prevalent in computer vision[1]



- More difficult for natural language
  - Discrete vocabulary
  - NMT sensitive to arbitrary noise

[1] image source:Medium

Existing Strategies

Language Technologies Institute

Word replacement

wie geht es dir ? ⟶ How are you ?

Existing Strategies

Language Technologies Institute

Word replacement

wie geht es dir ? ⟶ How are you ?

- Dictionary [Fadaee et al., 2017]
  wie geht es Tom ? ⟶ How is Tom ?

Existing Strategies

Word replacement

wie geht es dir ? ⟶ How are you ?

- Dictionary [Fadaee et al., 2017]
  wie geht es Tom ? ⟶ How is Tom ?
- Word dropout [Sennrich et al., 2016a]
  wie geht es NULL ? ⟶ How are you ?

Existing Strategies

Word replacement

wie geht es dir ? ⟶ How are you ?

- Dictionary [Fadaee et al., 2017]
  wie geht es Tom ? ⟶ How is Tom ?
- Word dropout [Sennrich et al., 2016a]
  wie geht es NULL ? ⟶ How are you ?
- Reward Augmented Maximum Likelihood (RAML) [Norouzi et al., 2016]
  wie geht es dir ? ⟶ How are hello ?

Existing Strategies

Language Technologies Institute

Word replacement

wie geht es dir ? ⟶ How are you ?

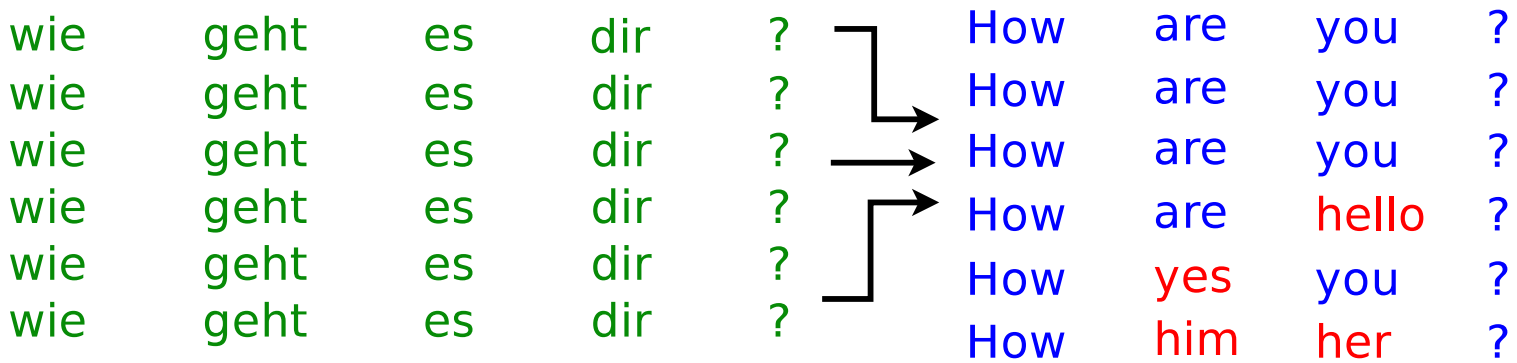
- Dictionary [Fadaee et al., 2017]
  wie geht es Tom ? ⟶ How is Tom ?
- Word dropout [Sennrich et al., 2016a]
  wie geht es NULL ? ⟶ How are you ?
- Reward Augmented Maximum Likelihood (RAML) [Norouzi et al., 2016]
  wie geht es dir ? ⟶ How are hello ?

→ Can we characterize all of the related approaches together?

Existing Strategies: RAML

RAML [Norouzi et al., 2016]

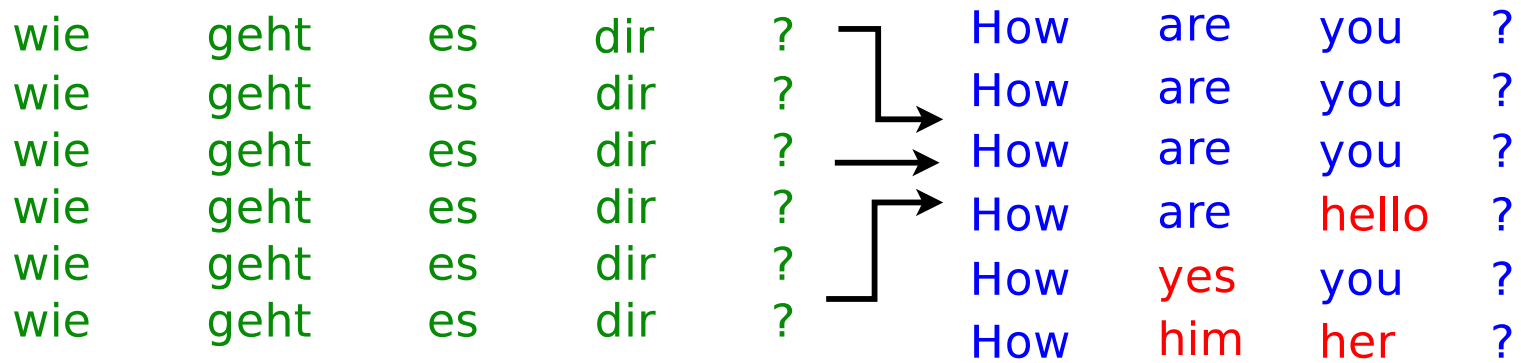
- Motivation: NMT relies on imperfect partial translation at test time, but trained only on gold standard target



Existing Strategies: RAML

RAML [Norouzi et al., 2016]

- Motivation: NMT relies on imperfect partial translation at test time, but trained only on gold standard target
- Solution: Sample corrupted target during training



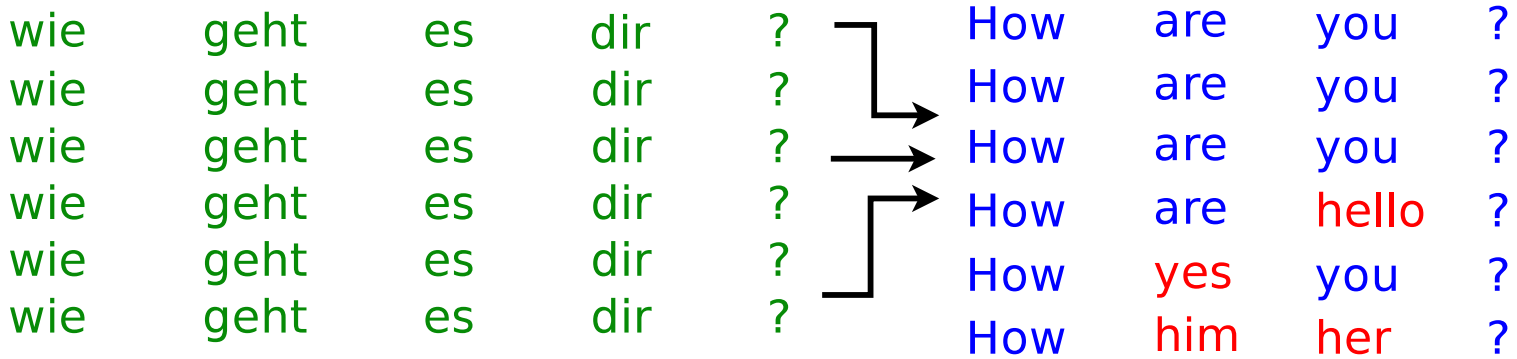
Existing Strategies: RAML

Language Technologies Institute

RAML [Norouzi et al., 2016]

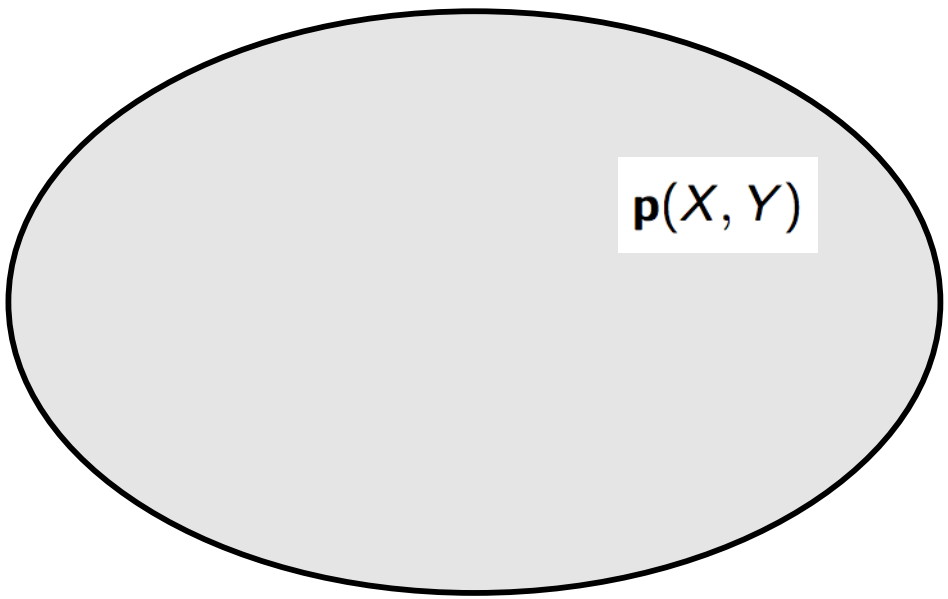
- Motivation: NMT relies on imperfect partial translation at test time, but trained only on gold standard target
- Solution: Sample corrupted target during training
- Gold target y, corrupted $\widehat{y}$, similarity measure r_y

$$\mathbf{q}^*(\widehat{y}|y,\tau) = \frac{\exp\{r_y(\widehat{y},y)/\tau\}}{\sum_{\widehat{y}'}\exp\{r_y(\widehat{y}',y)/\tau\}}$$



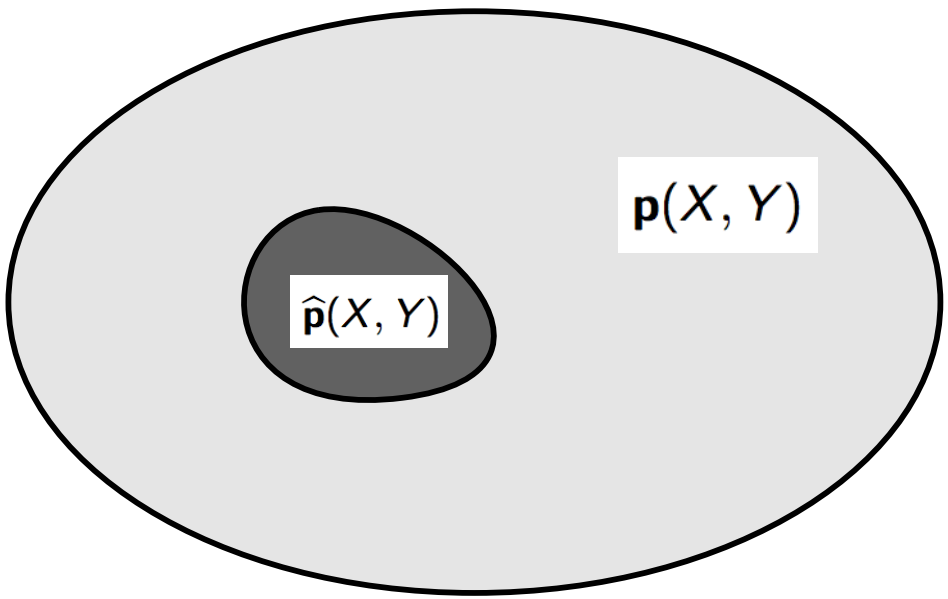
Formalize Data Augmentation

- **Real** data distribution: $x, y \sim \mathbf{p}(X, Y)$



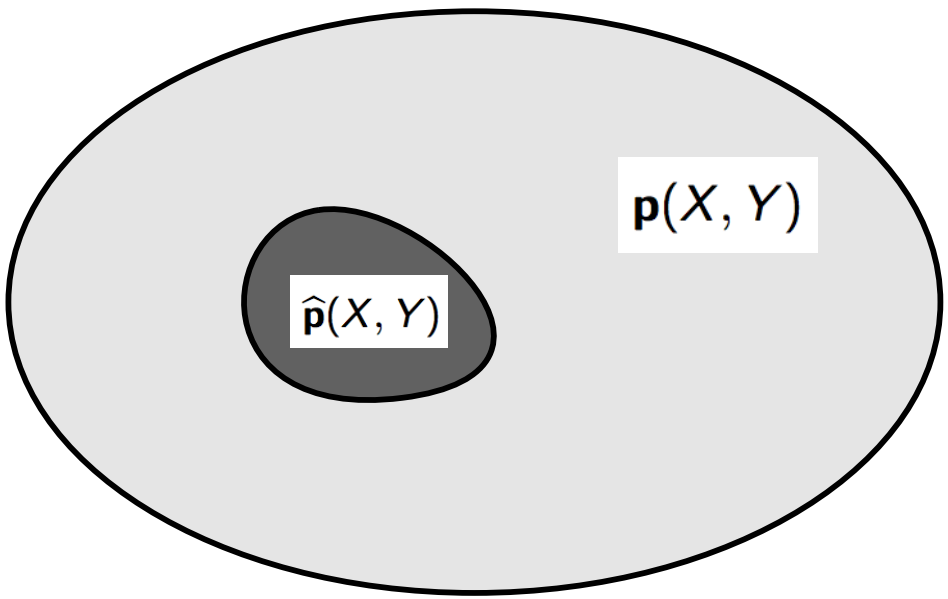
Formalize Data Augmentation

- **Real** data distribution: $x, y \sim \mathbf{p}(X, Y)$
- **Observed** data distribution: $x, y \sim \widehat{\mathbf{p}}(X, Y)$



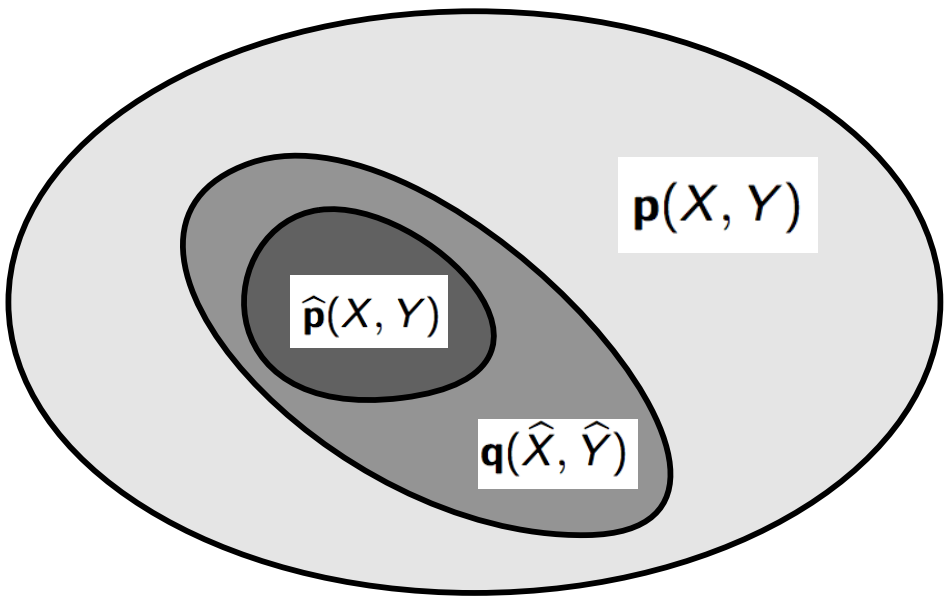
Formalize Data Augmentation

- **Real** data distribution: $x, y \sim \mathbf{p}(X, Y)$
- **Observed** data distribution: $x, y \sim \widehat{\mathbf{p}}(X, Y)$
  $\rightarrow$ Problem: $\mathbf{p}(X, Y)$ and $\widehat{\mathbf{p}}(X, Y)$ might have large discrepancy



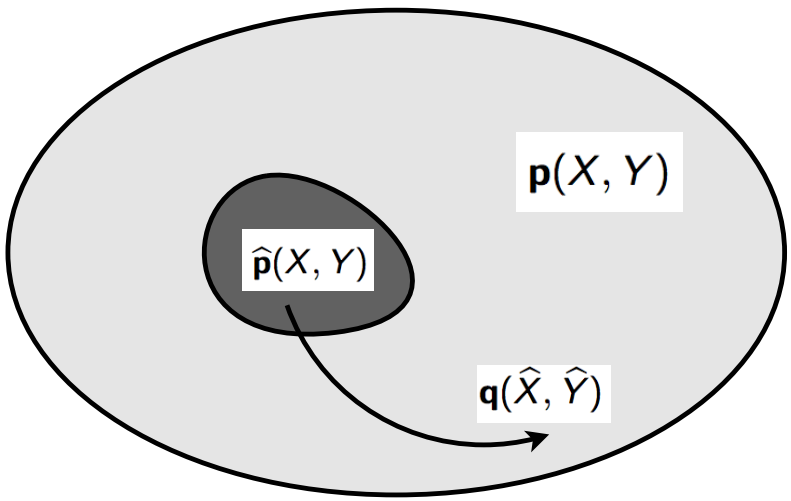
Formalize Data Augmentation

- **Real** data distribution: $x, y \sim \mathbf{p}(X, Y)$
- **Observed** data distribution: $x, y \sim \widehat{\mathbf{p}}(X, Y)$
  $\rightarrow$ Problem: $\mathbf{p}(X, Y)$ and $\widehat{\mathbf{p}}(X, Y)$ might have large discrepancy
- Data augmentation: $\widehat{x}, \widehat{y} \sim \mathbf{q}(\widehat{X}, \widehat{Y})$



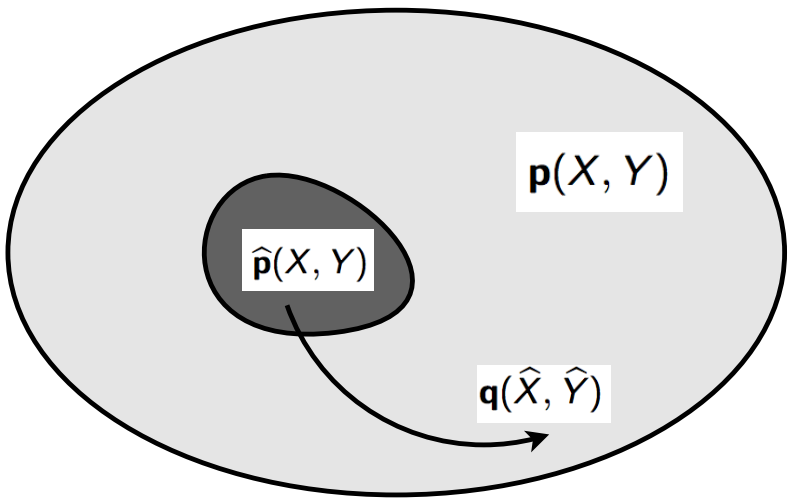
Design a good $\mathbf{q}(\widehat{X}, \widehat{Y})$

- **q**: function of **observed** (x, y)



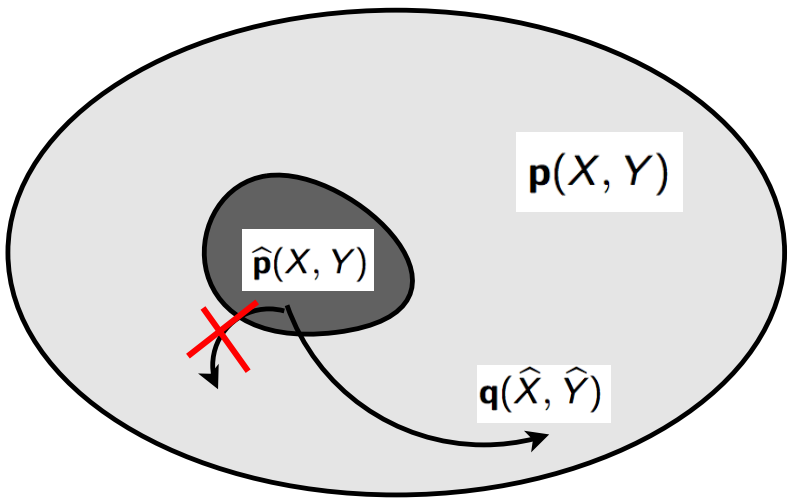
Design a good $\mathbf{q}(\widehat{X}, \widehat{Y})$

- **q**: function of **observed** (x, y)
- How should **q** approximate **p**?



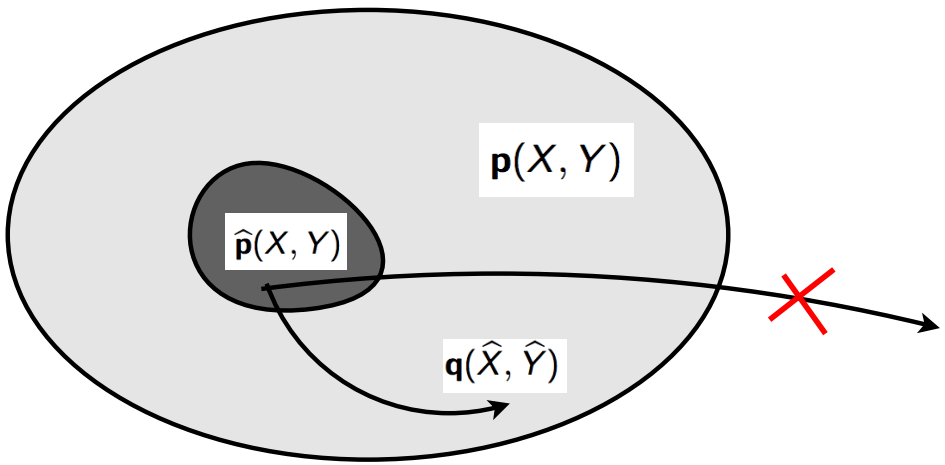
Design a good $\mathbf{q}(\widehat{X}, \widehat{Y})$

- **q**: function of **observed** (x, y)
- How should **q** approximate **p**?
  - **Diversity**: larger support with all valid data pairs (x, y)
    - ★ Entropy $\mathbb{H}\big[\mathbf{q}(\widehat{x}, \widehat{y}|x, y)\big]$ is large



Design a good $\mathbf{q}(\widehat{X}, \widehat{Y})$

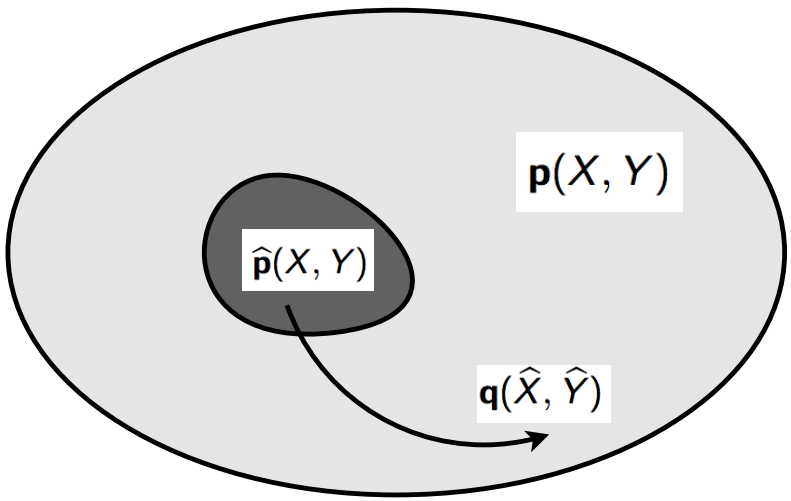
- **q**: function of **observed** (x, y)
- How should **q** approximate **p**?
  - **Diversity**: larger support with all valid data pairs (x, y)
    - ★ Entropy $\mathbb{H}\left[\mathbf{q}(\widehat{x}, \widehat{y}|x, y)\right]$ is large
  - **Smoothness**: probability of similar data pairs are similar
    - ★ **q** maximizes similarity measure $r_x(x, \widehat{x})$, $r_y(y, \widehat{y})$



Design a good $\mathbf{q}(\widehat{X}, \widehat{Y})$

- **q**: function of **observed** (x, y)
- How should **q** approximate **p**?
  - **Diversity**: larger support with all valid data pairs (x, y)
    - ★ Entropy $\mathbb{H}\left[\mathbf{q}(\widehat{x}, \widehat{y}|x, y)\right]$ is large
  - **Smoothness**: probability of similar data pairs are similar
    - ★ **q** maximizes similarity measure $r_x(x, \widehat{x})$, $r_y(y, \widehat{y})$
- τ: control effect of diversity; **q** should maximize

$$J(\mathbf{q}) = \tau \cdot \mathbb{H}\left[\mathbf{q}(\widehat{x}, \widehat{y}|x, y)\right] + \mathbb{E}_{\widehat{x},\widehat{y}\sim\mathbf{q}}\left[r_x(x, \widehat{x}) + r_y(y, \widehat{y})\right]$$



Mathematically Optimal $\mathbf{q}$

$$J(\mathbf{q}) = \tau \cdot \mathbb{H}\left[\mathbf{q}(\widehat{x}, \widehat{y}|x, y)\right] + \mathbb{E}_{\widehat{x},\widehat{y}\sim\mathbf{q}}\left[r_x(x, \widehat{x}) + r_y(y, \widehat{y})\right]$$

- Solve for the best $\mathbf{q}$

$$\mathbf{q}^*(\widehat{x}, \widehat{y}|x, y) = \frac{\exp\left\{s(\widehat{x}, \widehat{y}; x, y)/\tau\right\}}{\sum_{\widehat{x}',\widehat{y}'} \exp\left\{s(\widehat{x}', \widehat{y}'; x, y)/\tau\right\}}$$

Mathematically Optimal $\mathbf{q}$

$$J(\mathbf{q}) = \tau \cdot \mathbb{H}\left[\mathbf{q}(\widehat{x}, \widehat{y}|x, y)\right] + \mathbb{E}_{\widehat{x},\widehat{y}\sim\mathbf{q}}\left[r_x(x, \widehat{x}) + r_y(y, \widehat{y})\right]$$

- Solve for the best $\mathbf{q}$

$$\mathbf{q}^*(\widehat{x}, \widehat{y}|x, y) = \frac{\exp\left\{s(\widehat{x}, \widehat{y}; x, y)/\tau\right\}}{\sum_{\widehat{x}',\widehat{y}'} \exp\left\{s(\widehat{x}', \widehat{y}'; x, y)/\tau\right\}}$$

- Decompose x and y

$$\mathbf{q}^*(\widehat{x}, \widehat{y}|x, y) = \frac{\exp\left\{r_x(\widehat{x}, x)/\tau_x\right\}}{\sum_{\widehat{x}'} \exp\left\{r_x(\widehat{x}', x)/\tau_x\right\}} \times \frac{\exp\left\{r_y(\widehat{y}, y)/\tau_y\right\}}{\sum_{\widehat{y}'} \exp\left\{r_y(\widehat{y}', y)/\tau_y\right\}}$$

Mathematically Optimal $\mathbf{q}$

$$J(\mathbf{q}) = \tau \cdot \mathbb{H}\left[\mathbf{q}(\widehat{x}, \widehat{y}|x, y)\right] + \mathbb{E}_{\widehat{x},\widehat{y}\sim\mathbf{q}}\left[r_x(x, \widehat{x}) + r_y(y, \widehat{y})\right]$$

- Solve for the best $\mathbf{q}$

$$\mathbf{q}^*(\widehat{x}, \widehat{y}|x, y) = \frac{\exp\left\{s(\widehat{x}, \widehat{y}; x, y)/\tau\right\}}{\sum_{\widehat{x}',\widehat{y}'} \exp\left\{s(\widehat{x}', \widehat{y}'; x, y)/\tau\right\}}$$

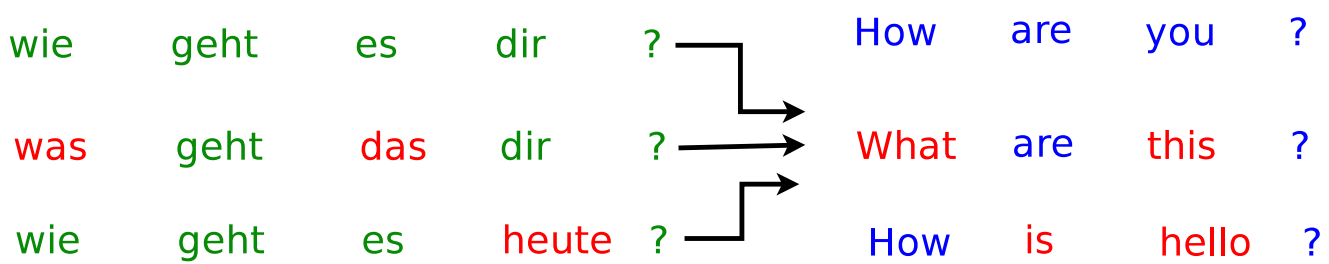
- Decompose x and y

$$\mathbf{q}^*(\widehat{x}, \widehat{y}|x, y) = \frac{\exp\left\{r_x(\widehat{x}, x)/\tau_x\right\}}{\sum_{\widehat{x}'} \exp\left\{r_x(\widehat{x}', x)/\tau_x\right\}} \times \frac{\exp\left\{r_y(\widehat{y}, y)/\tau_y\right\}}{\sum_{\widehat{y}'} \exp\left\{r_y(\widehat{y}', y)/\tau_y\right\}}$$

- Formulate existing methods
  - Dictionary: jointly on x and y, but deterministic and not diverse
  - Word dropout: **only** x side with null token
  - RAML: **only** y side

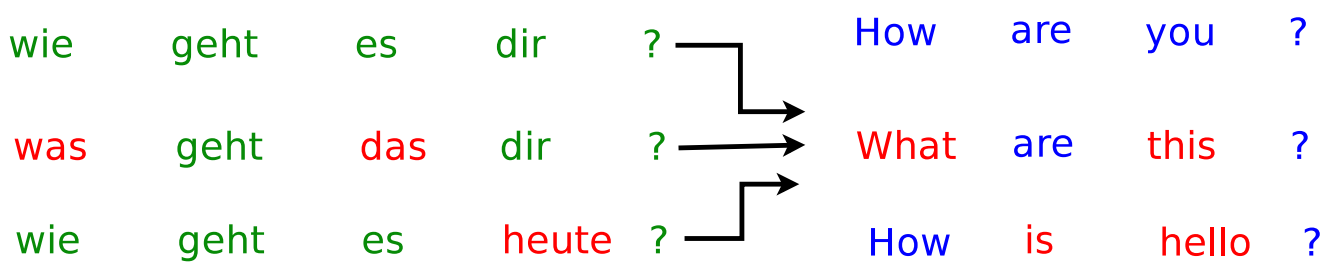
Formulate SwitchOut

- Augment **both** x and y!



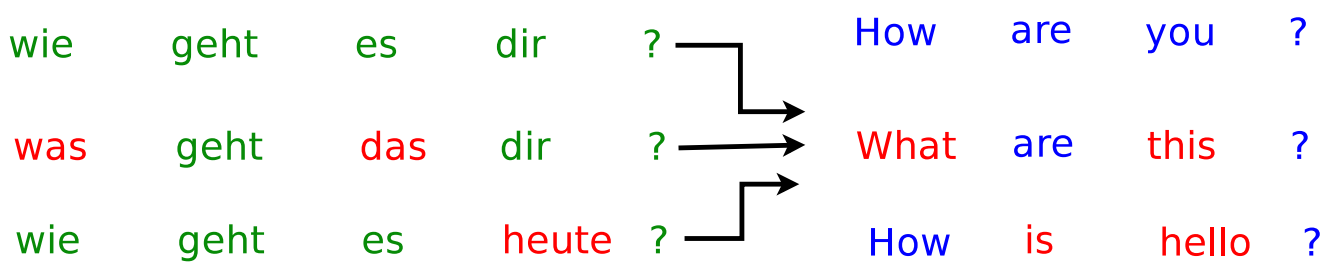
Formulate SwitchOut

- Augment **both** x and y!
- Sample for x, y **independently**



Formulate SwitchOut

- Augment **both** x and y!
- Sample for x, y **independently**
- Define $r_x(\widehat{x}, x)$ and $r_y(\widehat{y}, y)$
  - Negative Hamming Distance, following RAML



SwitchOut: Sample efficiently

Language Technologies Institute

Given a sentence $s = \{s_1, s_2, ...s_{|s|}\}$

1. How many words to corrupt?
   Assumption: only one token for swapping.

$$P(n) \propto \exp(-n)/\tau$$

SwitchOut: Sample efficiently

Language Technologies Institute

Given a sentence $s = \{s_1, s_2, ...s_{|s|}\}$

1. How many words to corrupt?
   Assumption: only one token for swapping.

$$P(n) \propto \exp(-n)/\tau$$

2. What is the corrupted sentence?

$$P(\text{randomly swap } s_i \text{ by another word}) = \frac{n}{|s|}$$

See Appendix: Efficient batch implementation in PyTorch and Tensorflow

Experiments

- Datasets
  - en-vi: IWSLT 2015
  - de-en: IWSLT 2016
  - en-de: WMT 2015
- Models
  - Transformer model
  - Word-based, standard preprocessing

Results: RAML and word dropout

Language Technologies Institute

Method		en-de	de-en	en-vi
src	**trg**			
N/A	N/A	21.73	29.81	27.97
WordDropout	N/A	20.63	**29.97**	28.56
SwitchOut	N/A	**22.78**†	29.94	**28.67**†
N/A	RAML	22.83	30.66	28.88
WordDropout	RAML	20.69	30.79	28.86
SwitchOut	RAML	**23.13**†	**30.98**†	**29.09**

Results: RAML and word dropout

Language Technologies Institute

- SwitchOut on source > word dropout

Method		en-de	de-en	en-vi
src	**trg**			
N/A	N/A	21.73	29.81	27.97
WordDropout	N/A	20.63	**29.97**	28.56
SwitchOut	N/A	**22.78**†	29.94	**28.67**†
N/A	RAML	22.83	30.66	28.88
WordDropout	RAML	20.69	30.79	28.86
SwitchOut	RAML	**23.13**†	**30.98**†	**29.09**

Results: RAML and word dropout

Language Technologies Institute

- SwitchOut on source > word dropout
- SwitchOut on source and target > RAML

Method		en-de	de-en	en-vi
src	**trg**			
N/A	N/A	21.73	29.81	27.97
WordDropout	N/A	20.63	**29.97**	28.56
SwitchOut	N/A	**22.78**†	29.94	**28.67**†
N/A	RAML	22.83	30.66	28.88
WordDropout	RAML	20.69	30.79	28.86
SwitchOut	RAML	**23.13**†	**30.98**†	**29.09**

Where does SwitchOut help?

- More gain for sentences more different from training data

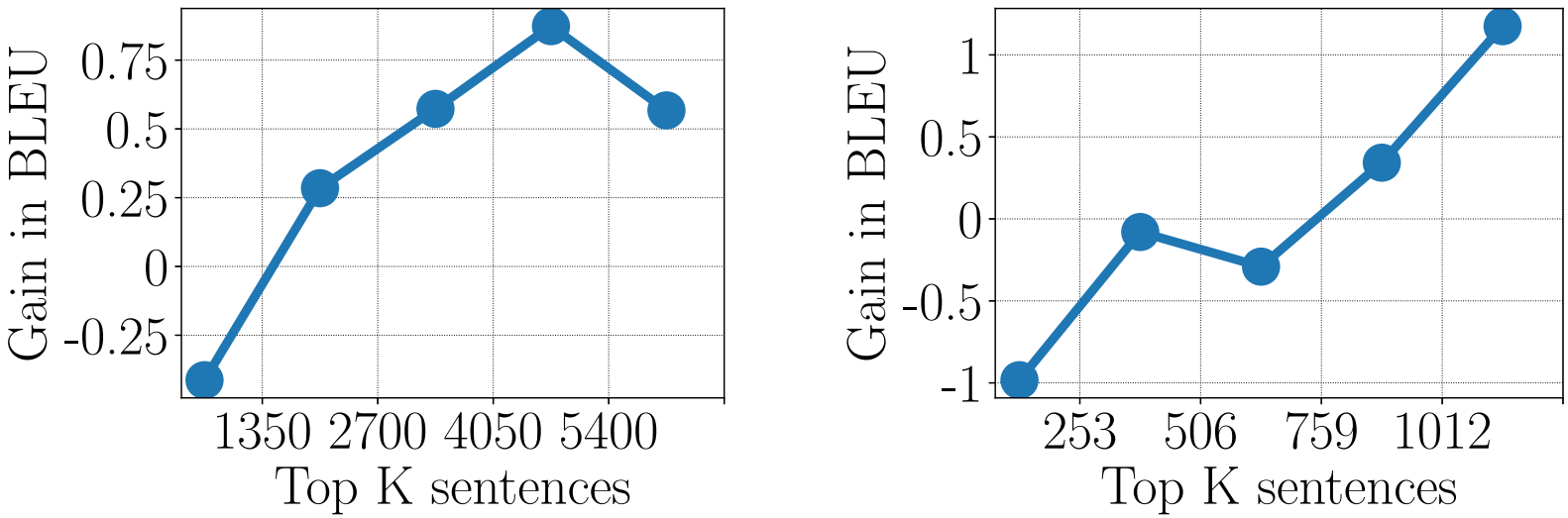


Figure: *Left*: IWSLT 16 de-en. *Right*: IWSLT 15 en-vi.

Final Thoughts

- SwitchOut sampling is efficient and easy-to-use

Final Thoughts

- SwitchOut sampling is efficient and easy-to-use
- Work with any NMT architecture

Final Thoughts

Language Technologies Institute

- SwitchOut sampling is efficient and easy-to-use
- Work with any NMT architecture
- Formulation of data augmentation encompasses existing works and inspires future direction

Final Thoughts

Language Technologies Institute

- SwitchOut sampling is efficient and easy-to-use
- Work with any NMT architecture
- Formulation of data augmentation encompasses existing works and inspires future direction

Thanks a lot for listening! Questions?

References

- Norouzi et al. (2016) Reward Augmented Maximum Likelihood for Neural Structured Prediction. In NIPS.
- Sennrich et al. (2016a) Edinburgh neural machine translation systems for wmt 16. In WMT.
- Sennrich et al. (2016b) Improving neural machine translation models with monolingual data. In ACL.
- Currey et al. (2017) Copied Monolingual Data Improves Low-Resource Neural Machine Translation. In WMT.
- Fadaee et al. (2017) Data Augmentation for Low-Resource Neural Machine Translation. In ACL.

Results: Back Translation (WMT en-de 2015)

- SwitchOut $>$ Back Translation

Method	**en-de**
Transformer	21.73
+SwitchOut	22.78
+BT	21.82

Results: Back Translation (WMT en-de 2015)

- SwitchOut > Back Translation
- Switchout + RAML + back translate wins

Method	**en-de**
Transformer	21.73
+SwitchOut	22.78
+BT	21.82
+BT +RAML	21.53
+BT +SwitchOut	22.93
+BT +RAML +SwitchOut	**23.76**