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

Language Technologies Institute

- 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

Word replacement

Word replacement

- Dictionary [Fadaee et al., 2017]

Word replacement

- Dictionary [Fadaee et al., 2017]
- Word dropout [Sennrich et al., 2016a]

Word replacement

- Dictionary [Fadaee et al., 2017]
- Word dropout [Sennrich et al., 2016a]
- Reward Augmented Maximum Likelihood (RAML) [Norouzi et al., 2016]

Existing Strategies

Word replacement

- Dictionary [Fadaee et al., 2017]
- Word dropout [Sennrich et al., 2016a]
- Reward Augmented Maximum Likelihood (RAML) [Norouzi et al., 2016]

! 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

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

Existing Strategies: RAML

Language Technologies Institute

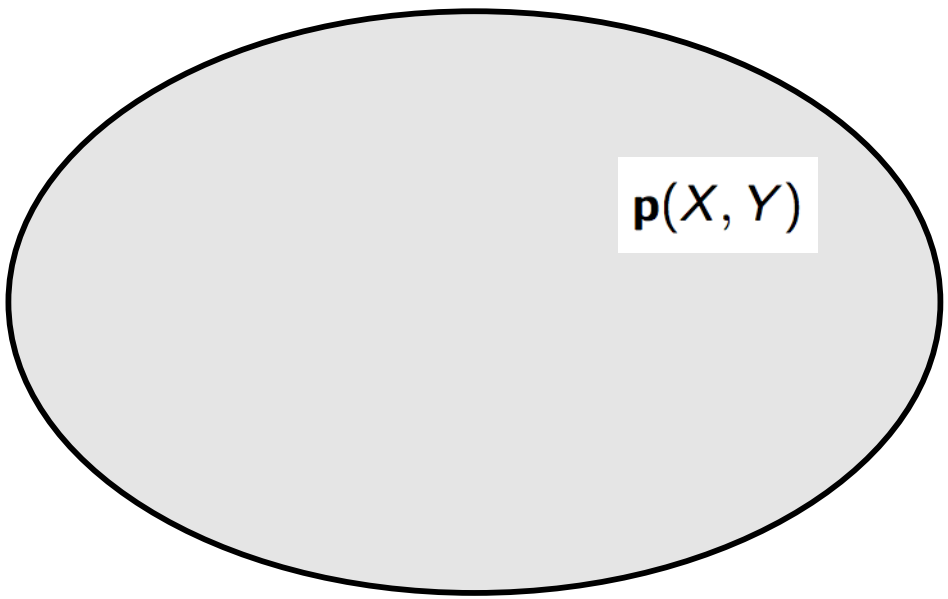
RAML [Norouzi et al., 2016]

- 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 $\hat{y}$, similarity measure r_y

$$\mathbf{q}^*(\hat{y}|y; \tau) = \frac{\exp f r_y(\hat{y}; y)/\tau g}{\sum_{\hat{y}'} \exp f r_y(\hat{y}'; y)/\tau g}$$

Formalize Data Augmentation

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



Formalize Data Augmentation

Real data distribution: $x; y \quad p(X; Y)$

Observed data distribution: $x; y \quad \hat{p}(X; Y)$

Formalize Data Augmentation

Real data distribution: $x; y \quad p(X; Y)$

Observed data distribution: $x; y \quad \hat{p}(X; Y)$

! Problem: $p(X; Y)$ and $\hat{p}(X; Y)$ might have large discrepancy

Formalize Data Augmentation

Real data distribution: $x; y \quad p(X; Y)$

Observed data distribution: $x; y \quad \hat{p}(X; Y)$

! Problem: $p(X; Y)$ and $\hat{p}(X; Y)$ might have large discrepancy

Data augmentation: $\hat{x}; \hat{y} \quad q(\hat{X}; \hat{Y})$

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

q: function of observed (x, y)

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

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

Design a good $q(\hat{p}; \hat{p})$

q: function of observed $(x; y)$

How should q approximate p?

- Diversity: larger support with all valid data pairs $x(, y)$
  - Entropy $H\ q(\hat{p}; \hat{p}|x; y)$ is large

Design a good $q(\hat{x}, \hat{y})$

q: function of observed (x, y)

How should q approximate p?

- Diversity: larger support with all valid data pairs $x(, y)$
  - Entropy $H\ q(\hat{x}, \hat{y}|x, y)$ is large
- Smoothness: probability of similar data pairs are similar
  - q maximizes similarity measure $r_x(x, \hat{x})$, $r_y(y, \hat{y})$

Design a good $q(\hat{x}, \hat{y})$

q: function of observed (x, y)

How should q approximate p?

- Diversity: larger support with all valid data pairs $x(, y)$
  - Entropy $H\ q(\hat{x}, \hat{y} | x, y)$ is large
- Smoothness: probability of similar data pairs are similar
  - q maximizes similarity measure $r_x(x, \hat{x})$, $r_y(y, \hat{y})$

: control effect of diversity; q should maximize

$$J(q) = H\ q(\hat{x}, \hat{y} | x, y) + E_{\hat{x}, \hat{y}\ q}\ r_x(x, \hat{x}) + r_y(y, \hat{y})$$

Mathematically Optimal q

$$J(q) = H\left(q(k_x, k_y \mid x, y)\right) + E_{k_x, k_y \sim q}\left[r_x(x, k_x) + r_y(y, k_y)\right]$$

Solve for the best q

$$q(k_x, k_y \mid x, y) = \frac{\exp\{s(k_x, k_y, x, y)/\tau\}}{\sum_{k_x', k_y'} \exp\{s(k_x', k_y', x, y)/\tau\}}$$

Mathematically Optimal q

$$J(q) = H\,q(\hat{x}, \hat{y} \mid x, y) + E_{\hat{x},\hat{y}\ q}\, r_x(x, \hat{x}) + r_y(y, \hat{y})$$

Solve for the best q

$$q^{\star}(\hat{x}, \hat{y} \mid x, y) = \frac{\exp\{-s(\hat{x}, \hat{y}, x, y)/\}}{\sum_{\hat{x}', \hat{y}'} \exp\{-s(\hat{x}', \hat{y}', x, y)/\}}$$

Decompose x and y

$$q^{\star}(\hat{x}, \hat{y} \mid x, y) = \frac{\exp\{-r_x(\hat{x}, x)/\ _x\}}{\sum_{\hat{x}'} \exp\{-r_x(\hat{x}', x)/\ _x\}} \quad \frac{\exp\{-r_y(\hat{y}, y)/\ _y\}}{\sum_{\hat{y}'} \exp\{-r_y(\hat{y}', y)/\ _y\}}$$

Mathematically Optimal q

$$J(q) = H\, q(\hat{x}, \hat{y} \mid x, y) + E_{\hat{x},\hat{y}\ q}\ r_x(x, \hat{x}) + r_y(y, \hat{y})$$

Solve for the best q

$$q^*(\hat{x}, \hat{y} \mid x, y) = \frac{\exp\{-s(\hat{x}, \hat{y}, x, y)/\ \}}{\sum_{\hat{x}', \hat{y}'} \exp\{-s(\hat{x}', \hat{y}', x, y)/\ \}}$$

Decompose x and y

$$q^*(\hat{x}, \hat{y} \mid x, y) = \frac{\exp\{-r_x(\hat{x}, x)/\ _x\}}{\sum_{\hat{x}'} \exp\{-r_x(\hat{x}', x)/\ _x\}} \quad \frac{\exp\{-r_y(\hat{y}, y)/\ _y\}}{\sum_{\hat{y}'} \exp\{-r_y(\hat{y}', y)/\ _y\}}$$

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 $r_x(\hat{x}, x)$ and $r_y(\hat{y}, y)$

- Negative Hamming Distance, following RAML

SwitchOut: Sample e ciently

Given a sentence $s = \{s_1; s_2; :::s_{|s|}\}$

How many words to corrupt?

Assumption: only one token for swapping.

$$P(n) / \exp(n)=$$

SwitchOut: Sample e ciently

Given a sentence $s = \{s_1; s_2; :::s_{|s|}\}$

How many words to corrupt?

Assumption: only one token for swapping.

$$P(n) / \exp(\ n)=$$

What is the corrupted sentence?

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

See Appendix: E cient batch implementation in PyTorch and Tensor ow

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

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

Results: RAML and word dropout

SwitchOut on source $\Rightarrow$ 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^{y}	29.94	28.67^{y}
N/A	RAML	22.83	30.66	28.88
WordDropout	RAML	20.69	30.79	28.86
SwitchOut	RAML	23.13^{y}	30.98^{y}	29.09

Results: RAML and word dropout

SwitchOut on source $\Rightarrow$ word dropout

SwitchOut on source and target $\Rightarrow$ RAML

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

Where does SwitchOut help?

- More gain for sentences more different from training data

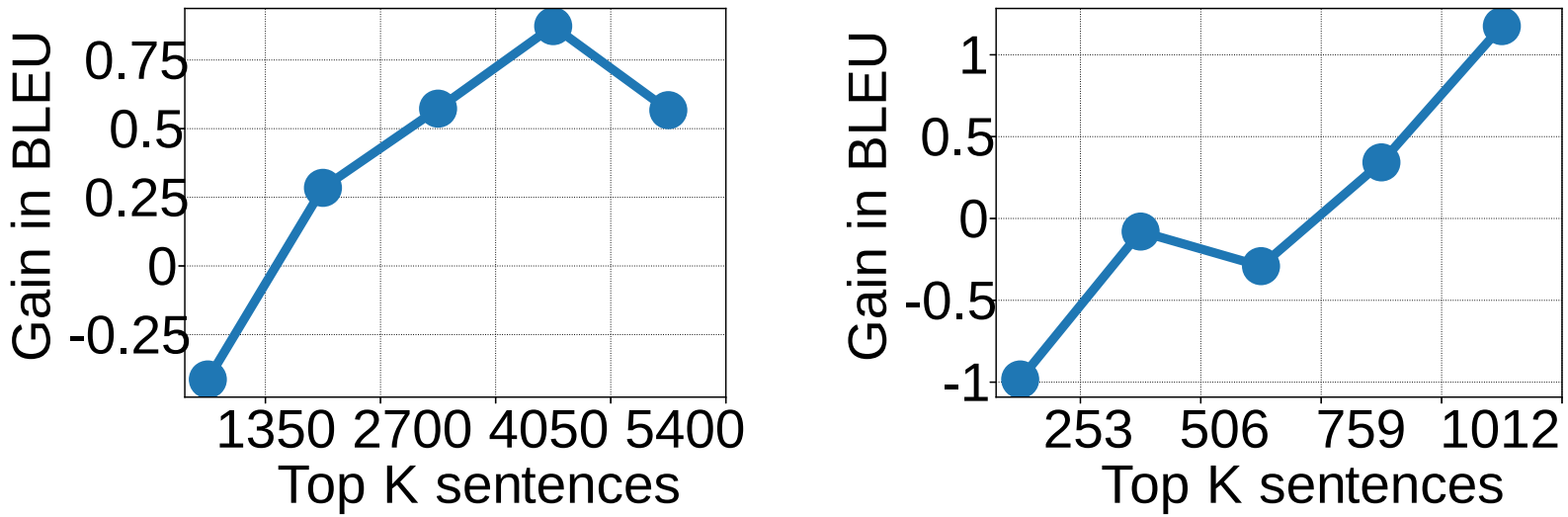


Figure: *Left*: IWSLT 16 de-en. *Right*: IWSLT 15 en-vi.

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**