達摩院
ALIBABA DAMO ACADEMY

Alibaba DAMO Academy for Discovery, Adventure, Momentum and Outlook

Learning to Generalize to More: Continuous Semantic Augmentation for Neural Machine Translation

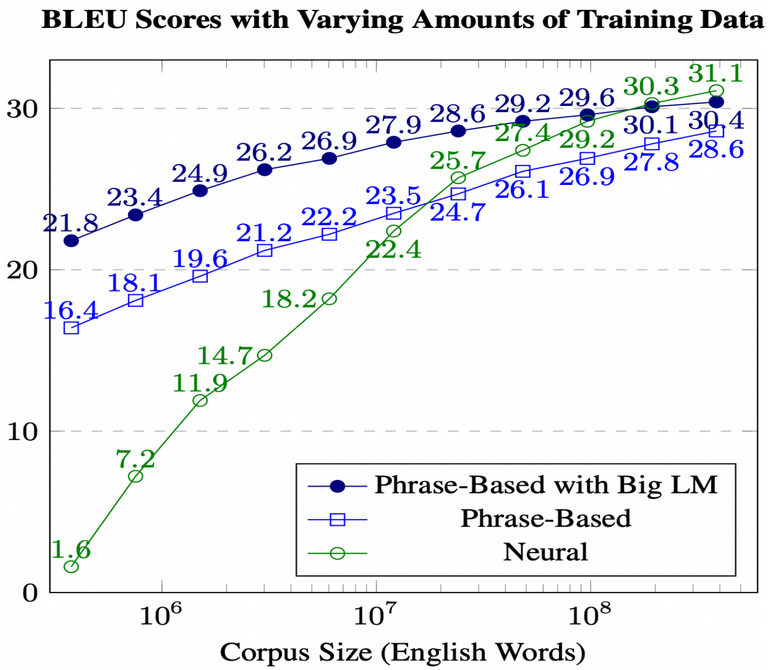
Xiangpeng Wei

Co-authors: Heng Yu, Yue Hu[†], Rongxiang Weng, Weihua Luo, Rong Jin

[†] From the University of Chinese Academy of Sciences (UCAS)

Neural Machine Translation

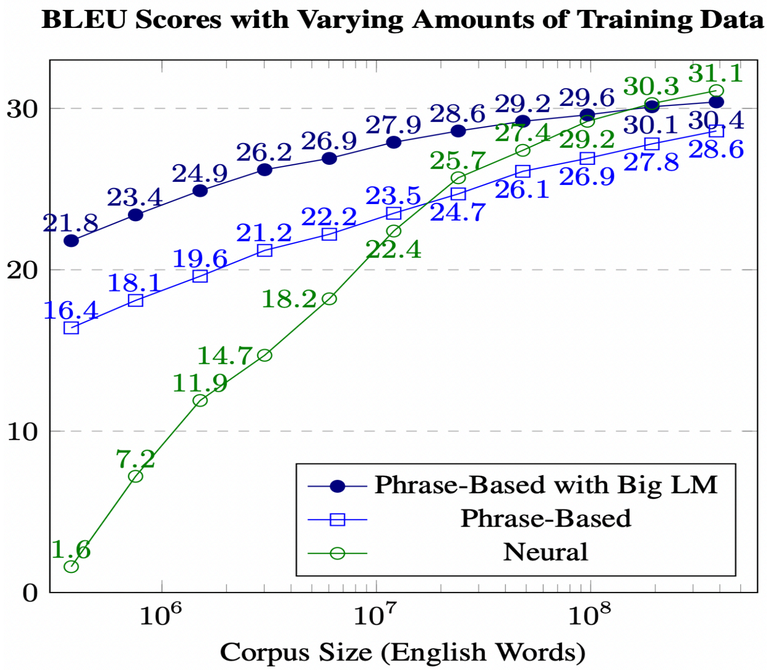
Neural Machine Translation (NMT) has developed into a dominant translation paradigm. However, its successes in many language pairs in the world depend heavily on large bitext corpora (primarily of human translations).



Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation.

Neural Machine Translation

Neural Machine Translation (NMT) has developed into a dominant translation paradigm. However, its successes in many language pairs in the world depend heavily on large bitext corpora (primarily of human translations).



NMT benefits more from the increasing amounts of training data, but it is hard to warmup with access to a few training instances.

Philipp Koehn and Rebecca Knowles. 2017. Six Challenges for Neural Machine Translation.

Data Augmentation

Increasing the amount of training data through automatic data augmentation (DA) techniques has wide applications in NLP.

Data Augmentation

Increasing the amount of training data through automatic data augmentation (DA) techniques has wide applications in NLP.

Previous Approaches:

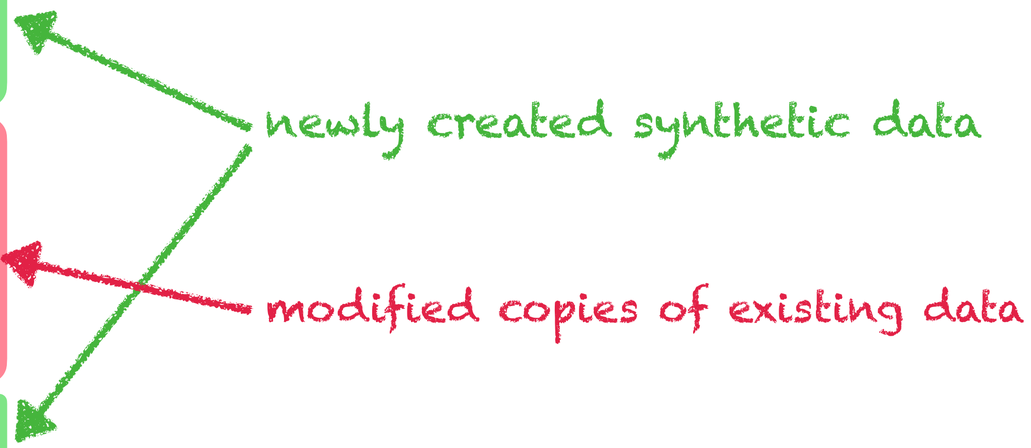
- Back-Translation (Sennrich et al., 2016; Ott et al., 2018)
- Self-Supervised Learning (Cheng et al., 2016)
- SwitchOut (Wang et al., 2018)
- Soft Contextual Data Augmentation (Gao et al., 2019)
- Robust NMT (Cheng et al., 2018, 2019, 2020)
- Data Diversification (Nguyen et al., 2020)
- …

Data Augmentation

Increasing the amount of training data through automatic data augmentation (DA) techniques has wide applications in NLP.

Previous Approaches:

- Back-Translation (Sennrich et al., 2016; Ott et al., 2018)
- Self-Supervised Learning (Cheng et al., 2016)
- SwitchOut (Wang et al., 2018)
- Soft Contextual Data Augmentation (Gao et al., 2019)
- Robust NMT (Cheng et al., 2018, 2019, 2020)
- Data Diversification (Nguyen et al., 2020)
- …



Previous Approaches: Back-Translation

Back-Translation (BT) makes use of the monolingual data on the target side to newly synthesize large scale pseudo parallel data, which is further combined with the real parallel corpus.

Sennrich et al. 2016. Improving neural machine translation models with monolingual data.

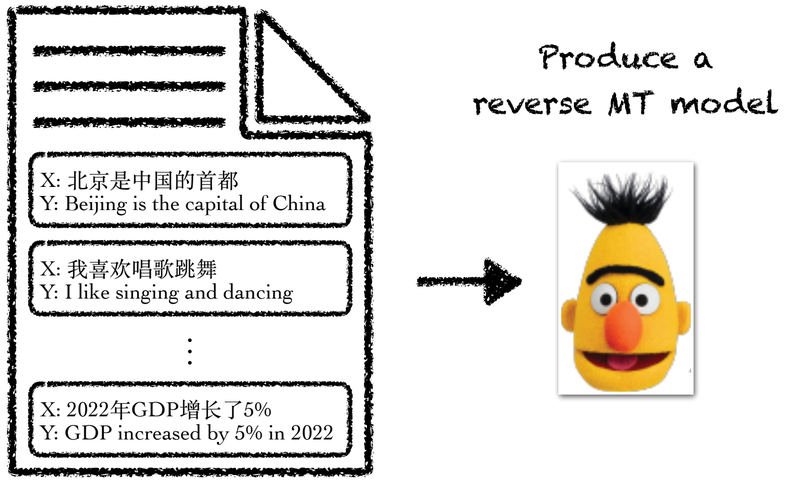
Previous Approaches: Back-Translation

Back-Translation (BT) makes use of the monolingual data on the target side to newly synthesize large scale pseudo parallel data, which is further combined with the real parallel corpus.

real bitext corpora

Previous Approaches: Back-Translation

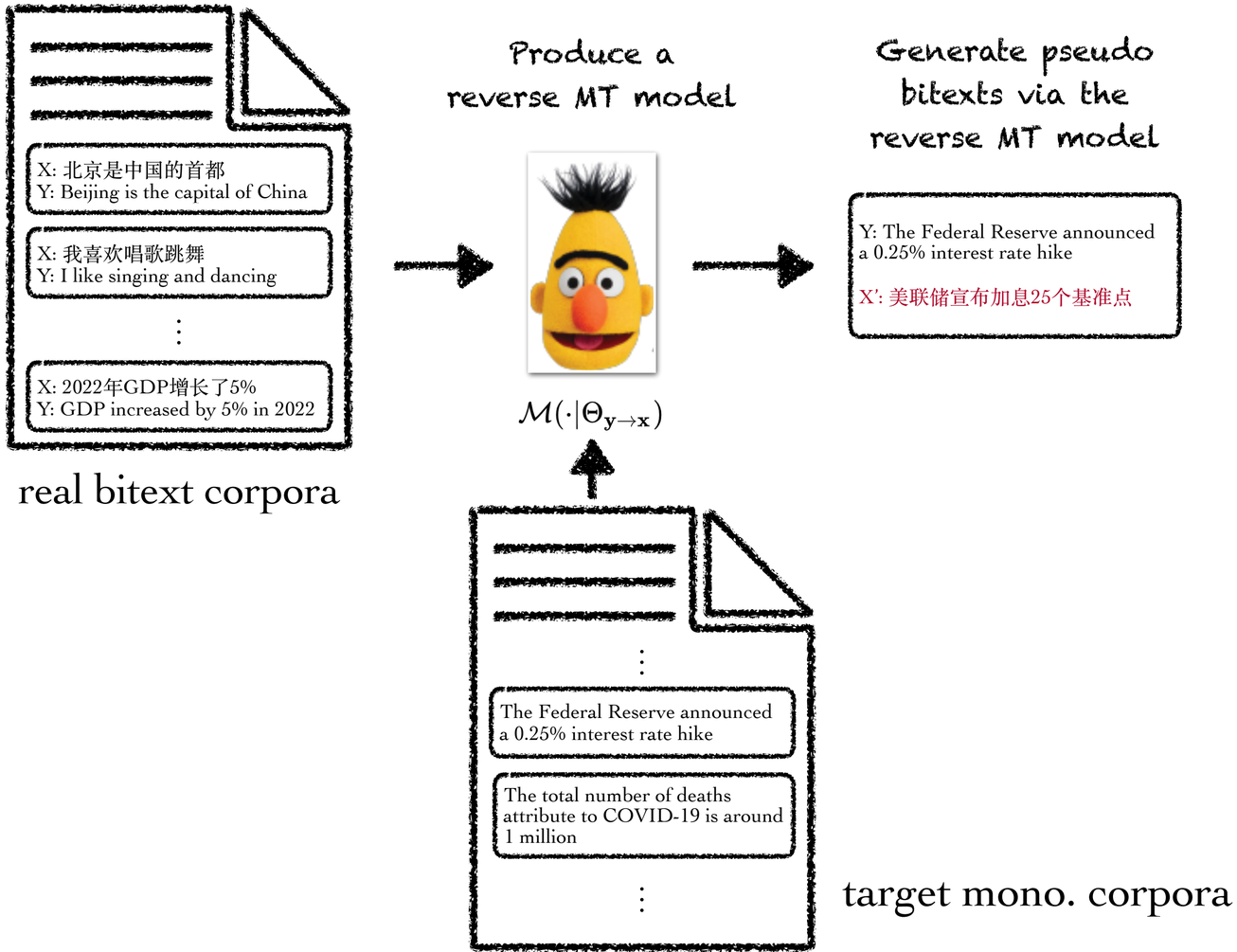
Back-Translation (BT) makes use of the monolingual data on the target side to newly synthesize large scale pseudo parallel data, which is further combined with the real parallel corpus.



real bitext corpora

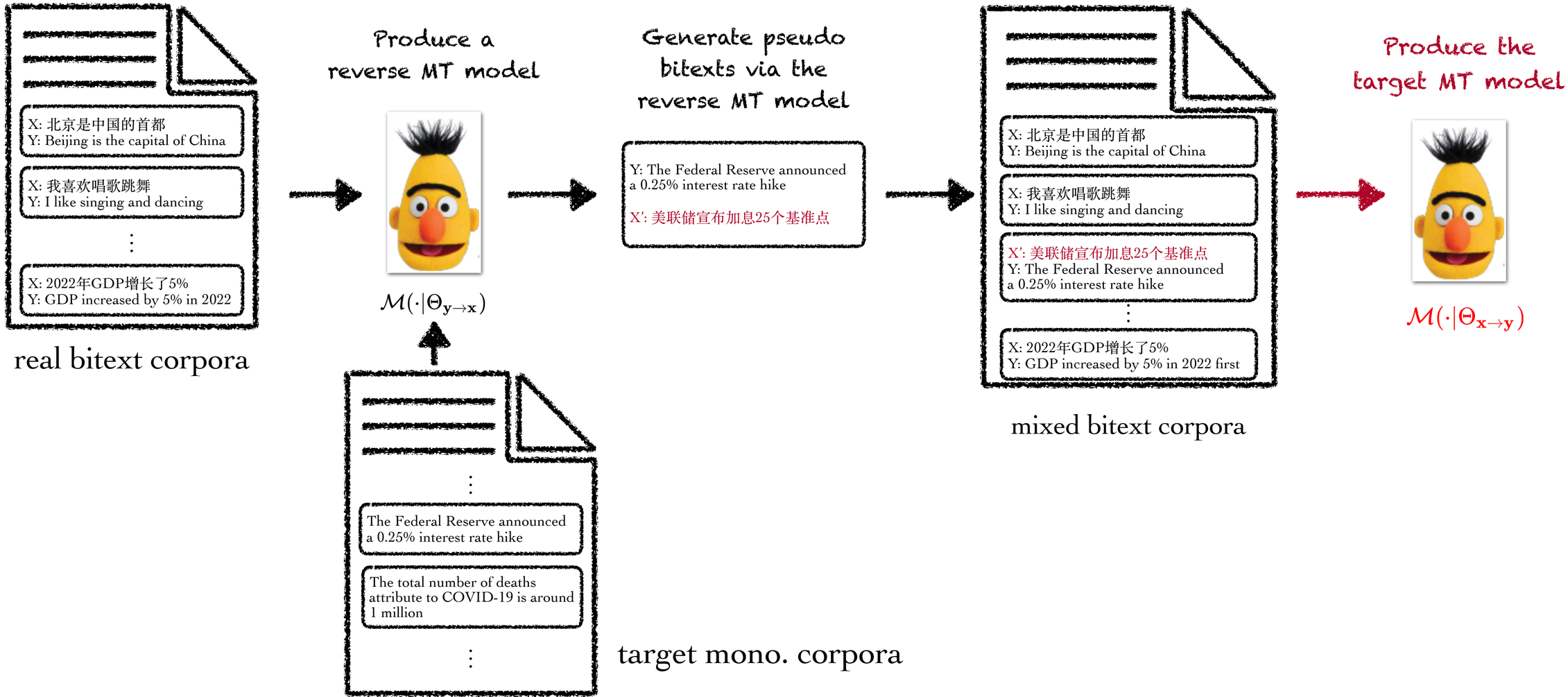
Previous Approaches: Back-Translation

Back-Translation (BT) makes use of the monolingual data on the target side to newly synthesize large scale pseudo parallel data, which is further combined with the real parallel corpus.



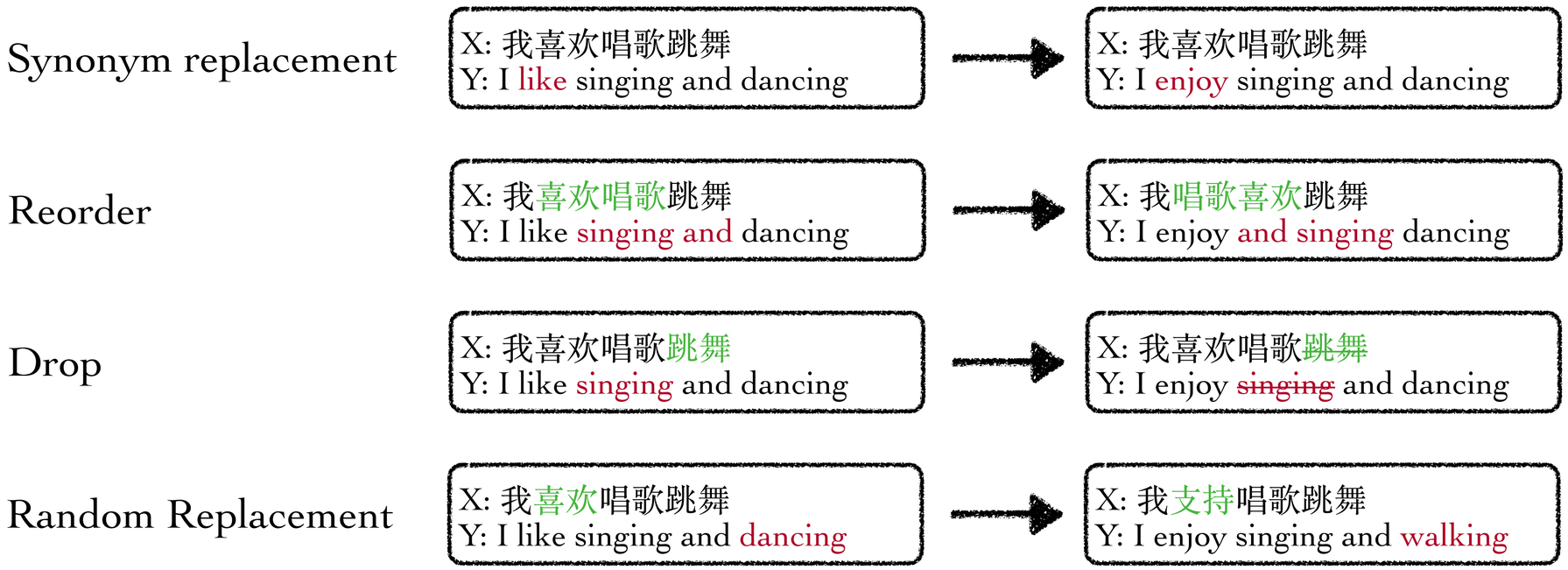
Previous Approaches: Back-Translation

Back-Translation (BT) makes use of the monolingual data on the target side to newly synthesize large scale pseudo parallel data, which is further combined with the real parallel corpus.



Previous Approaches: Adversarial Examples

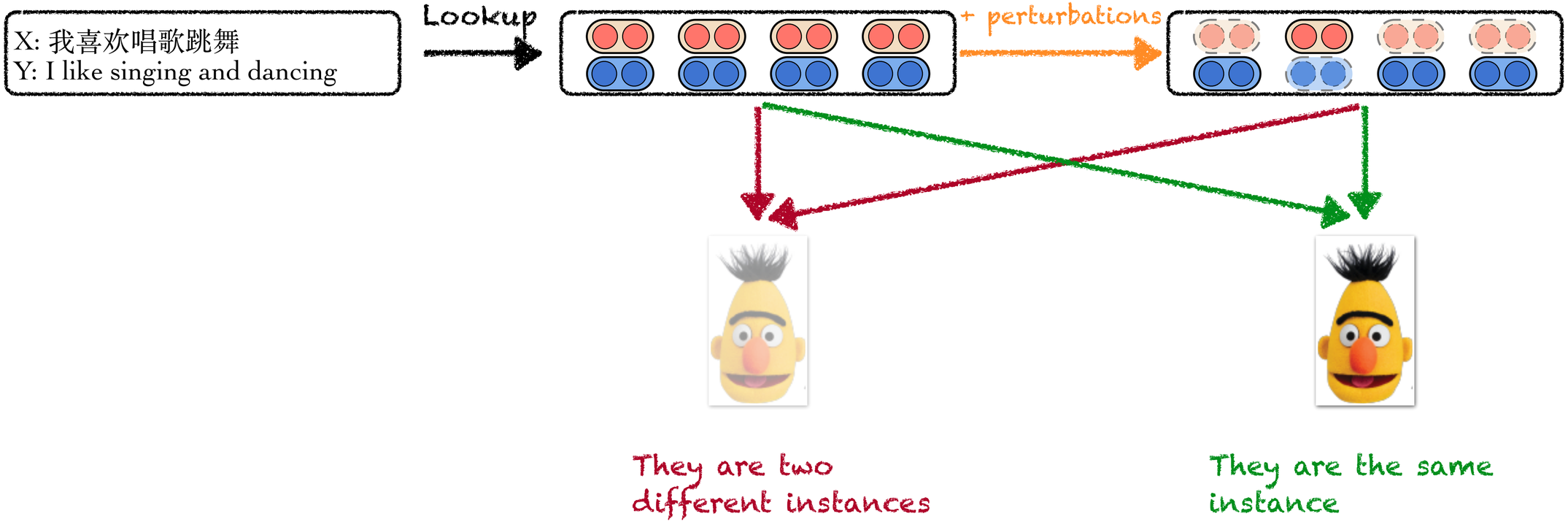
Producing modified copies of existing data via discrete manipulations such as adds, drops, reorders, and/or replaces words in original sentences.



Wang et al. 2018. SwitchOut: an efficient data augmentation algorithm for neural machine translation.

Previous Approaches: Adversarial Examples

Producing modified copies of existing data via discrete manipulations such as adds, drops, reorders, and/or replaces words in original sentences.



Cheng et al. 2018. Towards Robust Neural Machine Translation.

CsaNMT: Continuous Semantic Augmentation for NMT

- Learning a continuous semantic space to simulate the *real* distribution of semantic equivalence between source and target languages.
- Define a semantic vicinity for each observed training instance, which covers adequate variants of literal expression under the same meaning (many-to-many translation uncertainty).
- Optimizing the NMT model via vicinal risk minimization (i.e. maximize the log-likelihood of all training examples augmented with semantic vicinities).

Vicinal Risk Minimization

Formalizing data augmentation as enlarging the training set by drawing samples from a vicinity of existing training examples.

Vicinal Risk Minimization

Formalizing data augmentation as enlarging the training set by drawing samples from a vicinity of existing training examples.

The vanilla Empirical Risk Minimization (ERM):

transformation source target

$$R(f) = \int \ell(f(x), y) dP(x, y)$$

Vicinal Risk Minimization

Formalizing data augmentation as enlarging the training set by drawing samples from a vicinity of existing training examples.

The vanilla Empirical Risk Minimization (ERM):

transformation source target

$$R(f) = \int \ell(f(x), y) dP(x, y)$$

expectation loss of n observed samples

$$R_{erm}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Vicinal Risk Minimization

Formalizing data augmentation as enlarging the training set by drawing samples from a vicinity of existing training examples.

The vanilla Empirical Risk Minimization (ERM):

transformation source target

$$R(f) = \int \ell(f(x), y) dP(x, y)$$

expectation loss of n observed samples

$$R_{erm}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Vicinal Risk Minimization (VRM): an extension of ERM with a vicinity per sample

$$R_{vrm}(f) = \frac{1}{n} \sum_{i=1}^{n} \int \ell(f(x), y_i) dP_{x_i}(x)$$

the vicinity

Formalization of CsaNMT

$(\mathbf{x}, \mathbf{y}) \in (X, Y)$ denotes a pair of two sentences with the same meaning, i.e. a parallel sentence pair

Maximum Likelihood Estimation

training corpora parameters

$$J_{mle}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{C}}\big(\log P(\mathbf{y}|\mathbf{x};\Theta)\big)$$

decomposed to

$$J_{mle}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{C}}\Big[\sum_{t=1}^{T'} \log P(y_t|\mathbf{y}_{<t},\mathbf{x};\Theta)\Big]$$

Formalization of CsaNMT

$(\mathbf{x}, \mathbf{y}) \in (X, Y)$ denotes a pair of two sentences with the same meaning, i.e. a parallel sentence pair

Maximum Likelihood Estimation

training corpora parameters

$$J_{mle}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{C}}\big(\log P(\mathbf{y}|\mathbf{x};\Theta)\big)$$

decomposed to

$$J_{mle}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{C}}\Big[\sum_{t=1}^{T'} \log P(y_t|\mathbf{y}_{<t},\mathbf{x};\Theta)\Big]$$

Maximum Likelihood Estimation with Vicinities

$$J_{mle}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{C},\hat{r}^{(k)}\in\mathcal{R}}\big(\log P(\mathbf{y}|\mathbf{x},\hat{r}^{(k)};\Theta)\big)$$

Formalization of CsaNMT

$(\mathbf{x}, \mathbf{y}) \in (X, Y)$ denotes a pair of two sentences with the same meaning, i.e. a parallel sentence pair

Maximum Likelihood Estimation

training corpora parameters

$$J_{mle}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{C}}\big(\log P(\mathbf{y}|\mathbf{x};\Theta)\big)$$

decomposed to

$$J_{mle}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{C}}\Big[\sum_{t=1}^{T'} \log P(y_t|\mathbf{y}_{<t},\mathbf{x};\Theta)\Big]$$

Maximum Likelihood Estimation with Vicinities

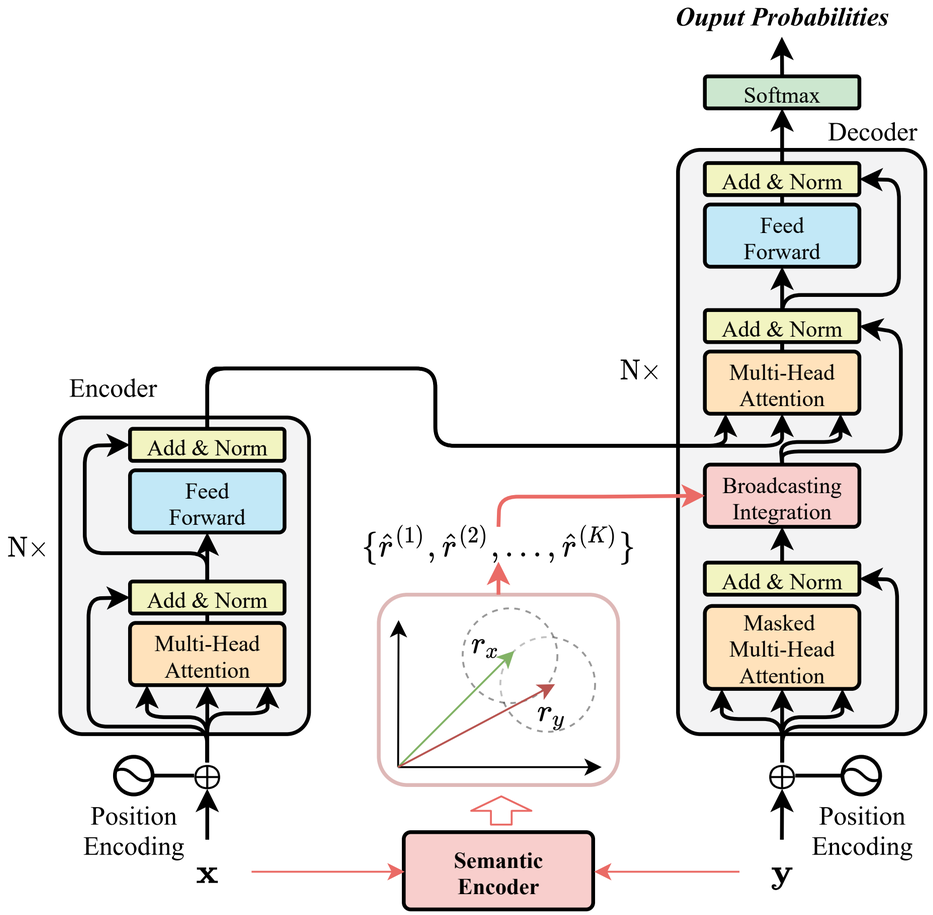
$$J_{mle}(\Theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{C},\hat{r}^{(k)}\in\mathcal{R}}\big(\log P(\mathbf{y}|\mathbf{x},\hat{r}^{(k)};\Theta)\big)$$

$$\mathcal{R} = \{\hat{r}^{(1)}, \hat{r}^{(2)}, \ldots, \hat{r}^{(K)}\}$$
$$\hat{r}^{(k)} \sim \nu(r_x, r_y)$$

a series of examples sampled from the vicinity

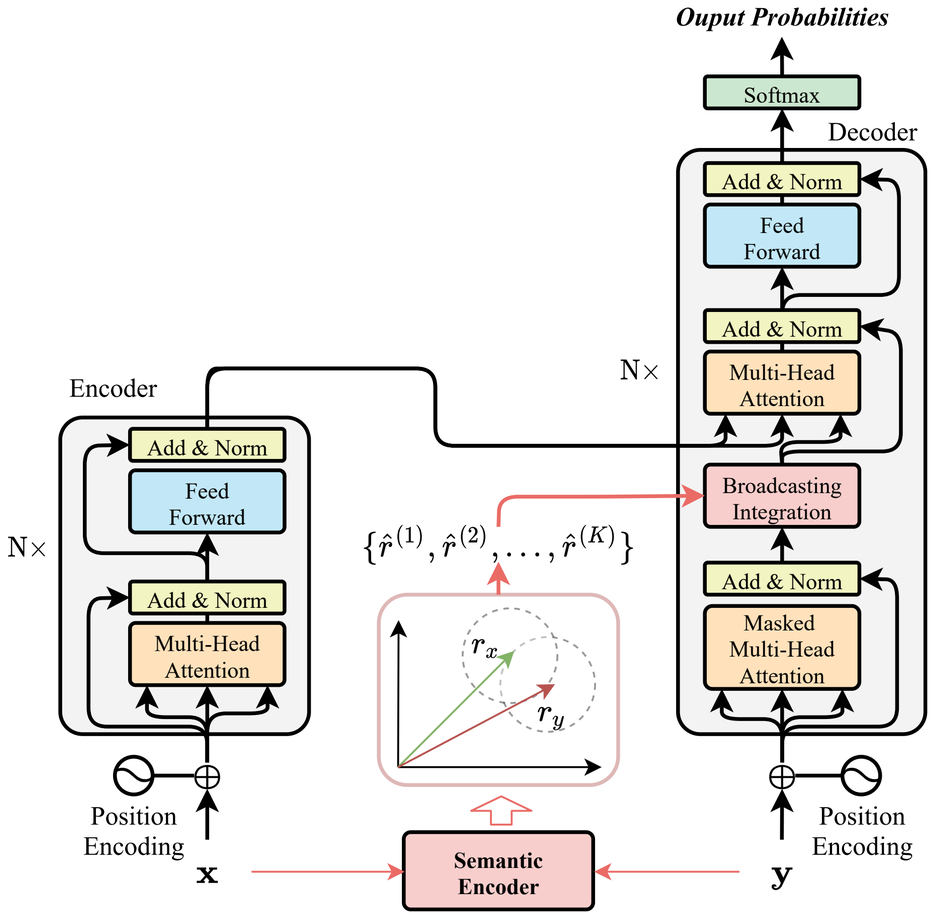
Graphical Illustration of CsaNMT

The model architecture consists of a vanilla encoder-decoder pipeline and an extra semantic encoder.



Graphical Illustration of CsANMT

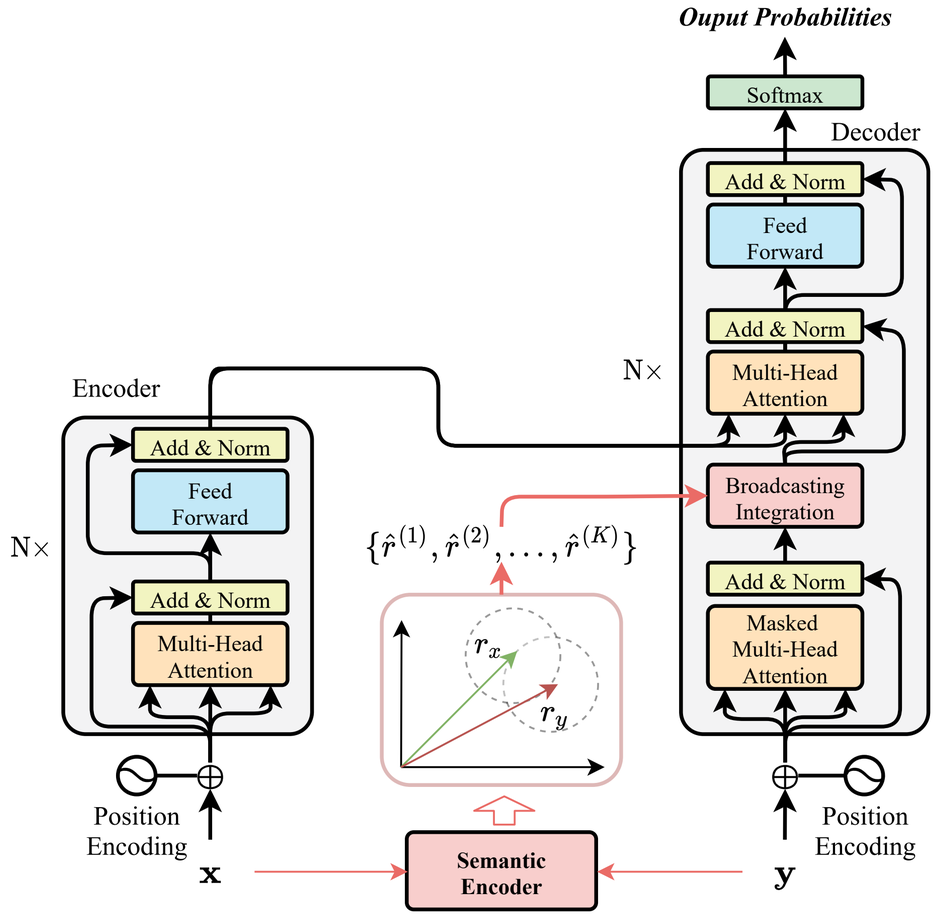
The model architecture consists of a vanilla encoder-decoder pipeline and an extra semantic encoder.



Definition 1. There is a universal semantic space among the source and the target languages for neural machine translation, which is established by a semantic encoder. It defines a forward function $\psi(\cdot\,;\Theta')$ to map discrete sentences into continuous vectors, that satisfies: $\forall(\mathbf{x},\mathbf{y}) \in (\mathcal{X},\mathcal{Y}) : r_x = r_y$. Besides, an adjacency semantic region $\nu(r_x, r_y)$ in the semantic space describes adequate variants of literal expression centered around each observed sentence pair $(\mathbf{x},\mathbf{y})$.

Graphical Illustration of CsaNMT

The model architecture consists of a vanilla encoder-decoder pipeline and an extra semantic encoder.



Definition 1. There is a universal semantic space among the source and the target languages for neural machine translation, which is established by a semantic encoder. It defines a forward function $\psi(\cdot\,;\Theta')$ to map discrete sentences into continuous vectors, that satisfies: $\forall(\mathbf{x},\mathbf{y}) \in (\mathcal{X},\mathcal{Y}) : r_x = r_y$. Besides, an adjacency semantic region $\nu(r_x, r_y)$ in the semantic space describes adequate variants of literal expression centered around each observed sentence pair $(\mathbf{x},\mathbf{y})$.

- How to optimize the extra semantic encoder $\psi(\cdot\,;\Theta')$ so that it produces a meaningful vicinity $\nu(r_x, r_y)$ for each training instance?
- How to obtain samples from the adjacency semantic region in an efficient and effective way?

Tangential Contrastive Learning

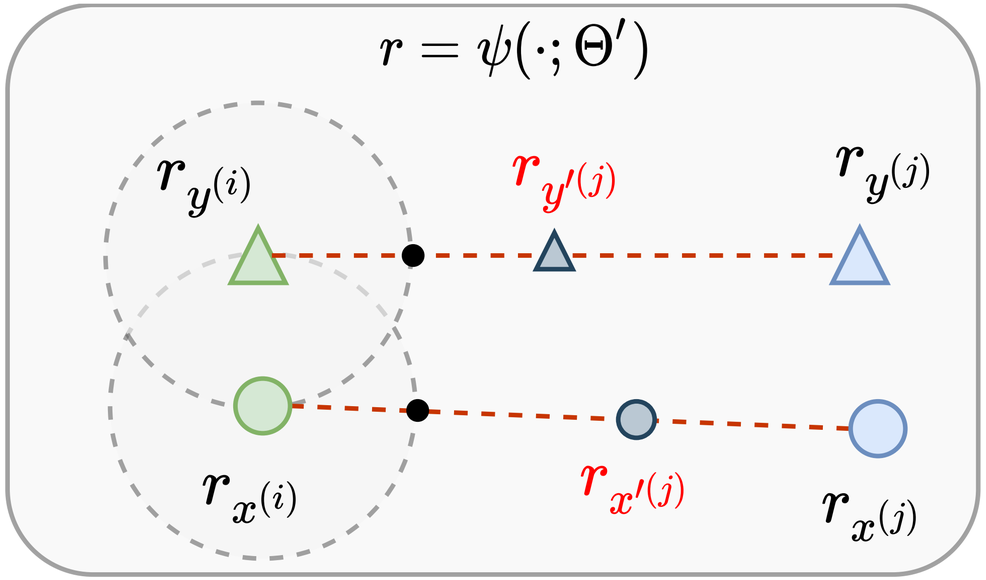
Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)

Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

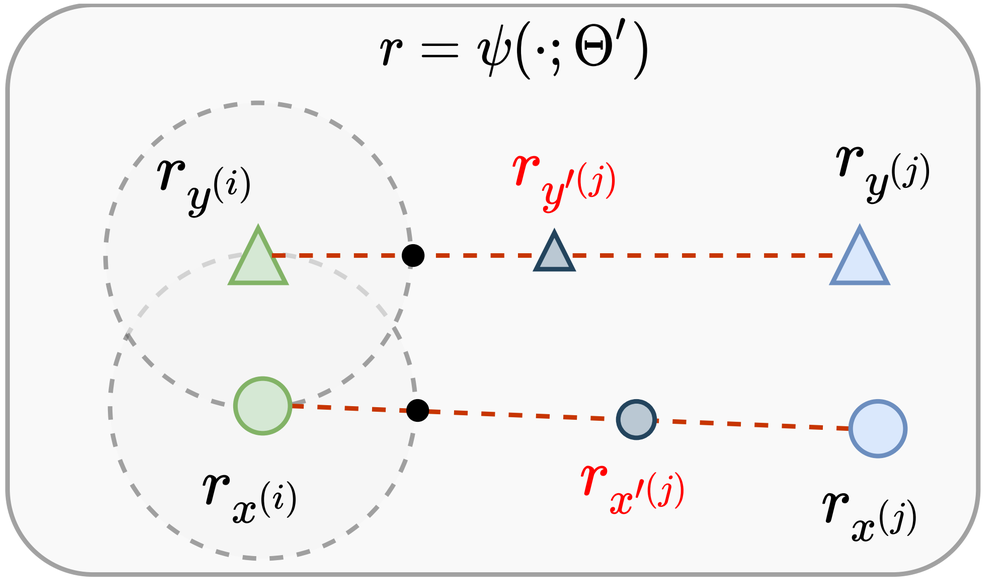
Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)



Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)

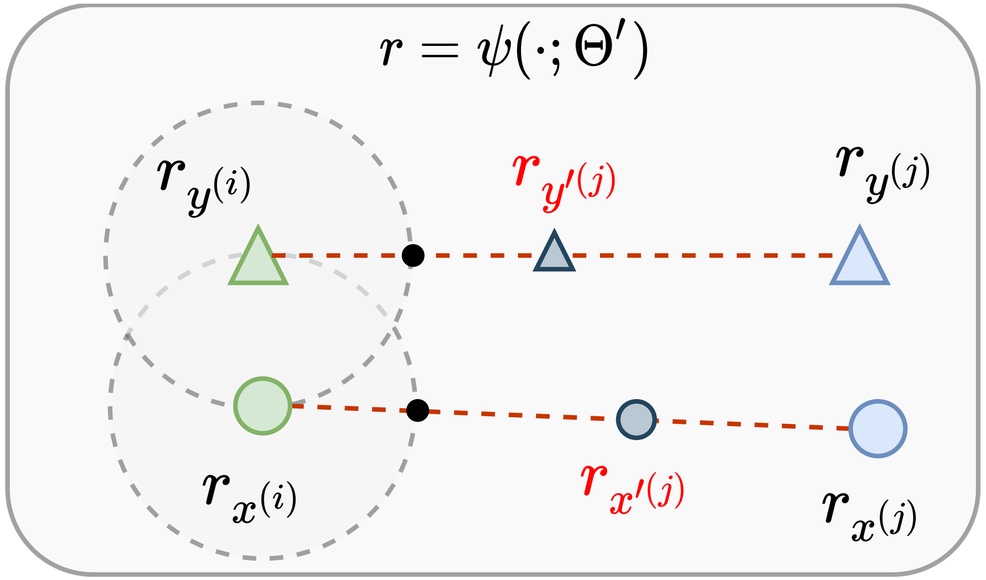


$\nu(\mathbf{r_x}, \mathbf{r_y})$: the union of two closed balls that are centered around r_x and r_y, the radius is $d = \| r_x - r_y \|_2$

Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)



$\nu(\mathbf{r_x}, \mathbf{r_y})$: the union of two closed balls that are centered around r_x and r_y, the radius is $d = \| r_x - r_y \|_2$

- i and j are two instances randomly sampled from the training corpora.

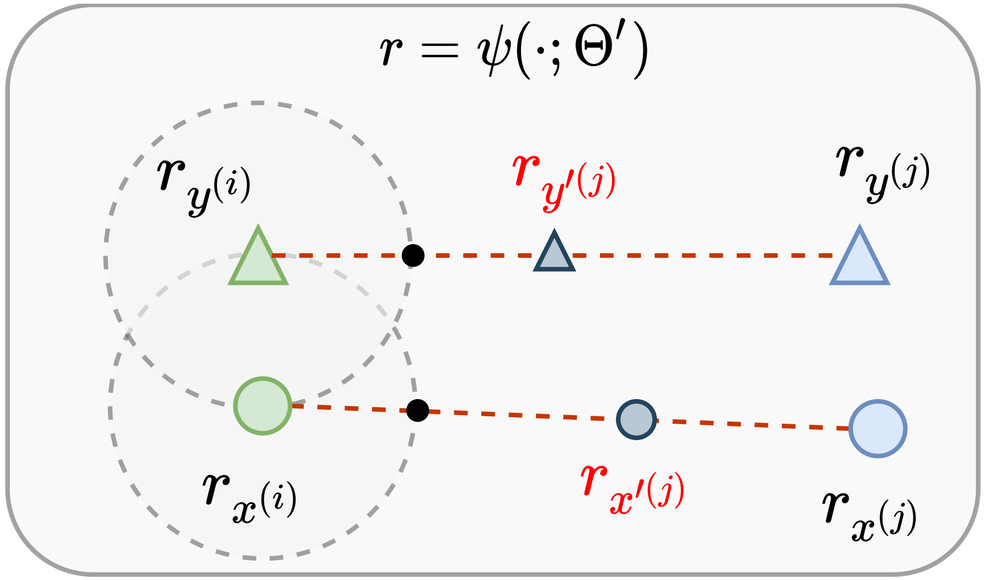
$$r_{x'(j)} = r_{x(i)} + \lambda_x (r_{x(j)} - r_{x(i)}), \lambda_x \in (\frac{d}{d'_x}, 1]$$

$$r_{y'(j)} = r_{y(i)} + \lambda_y (r_{y(j)} - r_{y(i)}), \lambda_y \in (\frac{d}{d'_y}, 1]$$

Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)



$\nu(\mathbf{r_x}, \mathbf{r_y})$: the union of two closed balls that are centered around r_x and r_y, the radius is $d = \| r_x - r_y \|_2$

- i and j are two instances randomly sampled from the training corpora.

$$r_{x'(j)} = r_{x(i)} + \lambda_x (r_{x(j)} - r_{x(i)}), \lambda_x \in (\frac{d}{d'_x}, 1]$$

$$r_{y'(j)} = r_{y(i)} + \lambda_y (r_{y(j)} - r_{y(i)}), \lambda_y \in (\frac{d}{d'_y}, 1]$$

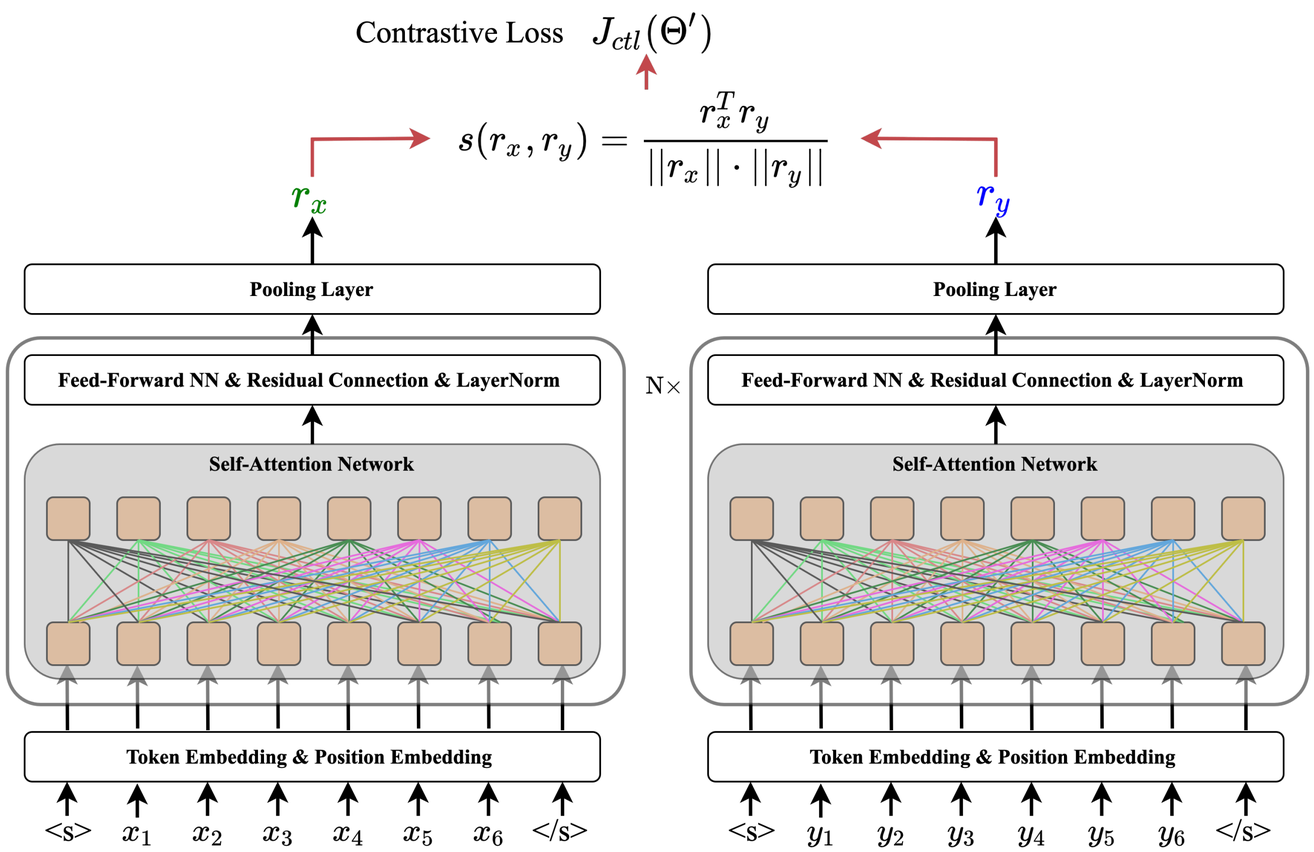
Adaptively adjusted factors during training

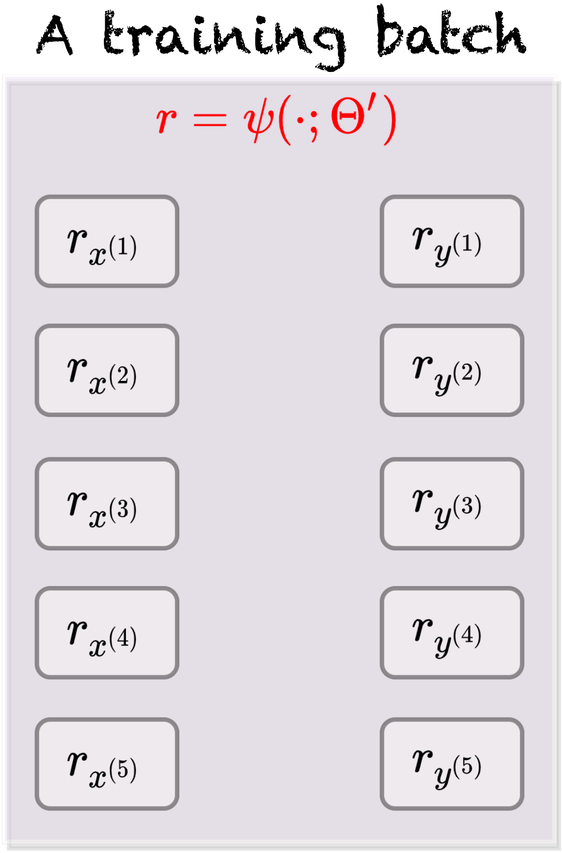
Reasonable Negatives are walking between tangent points and in-batch ones

Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)

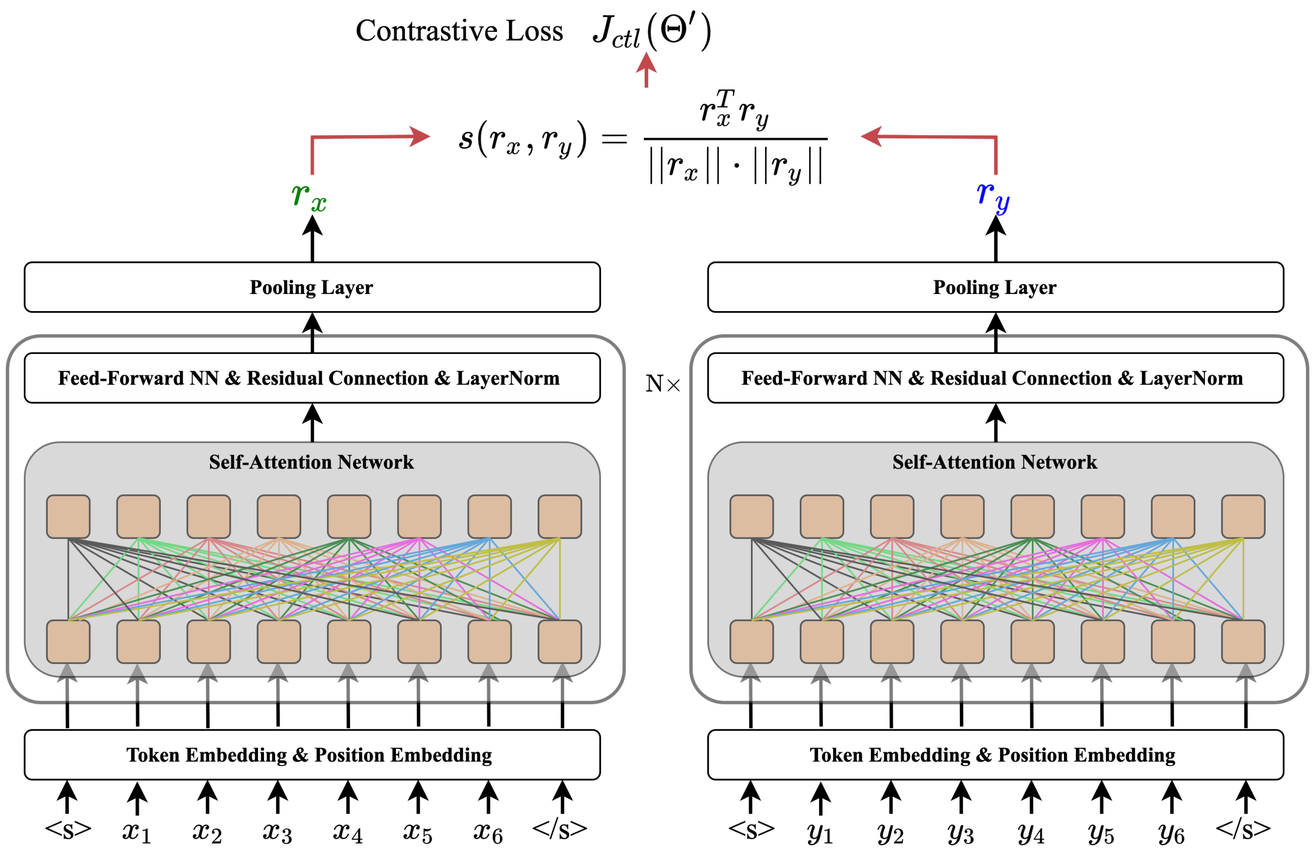


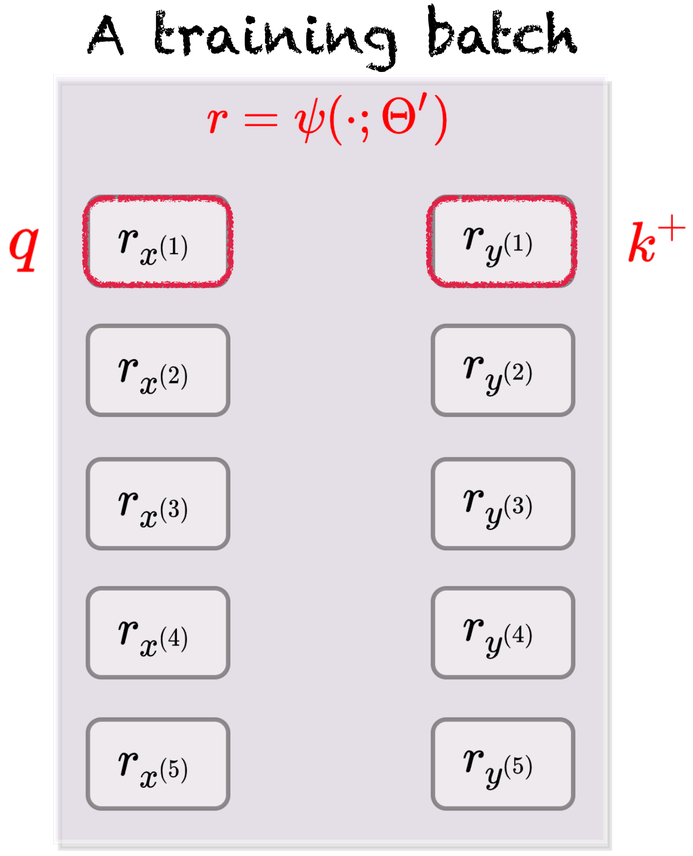


Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)

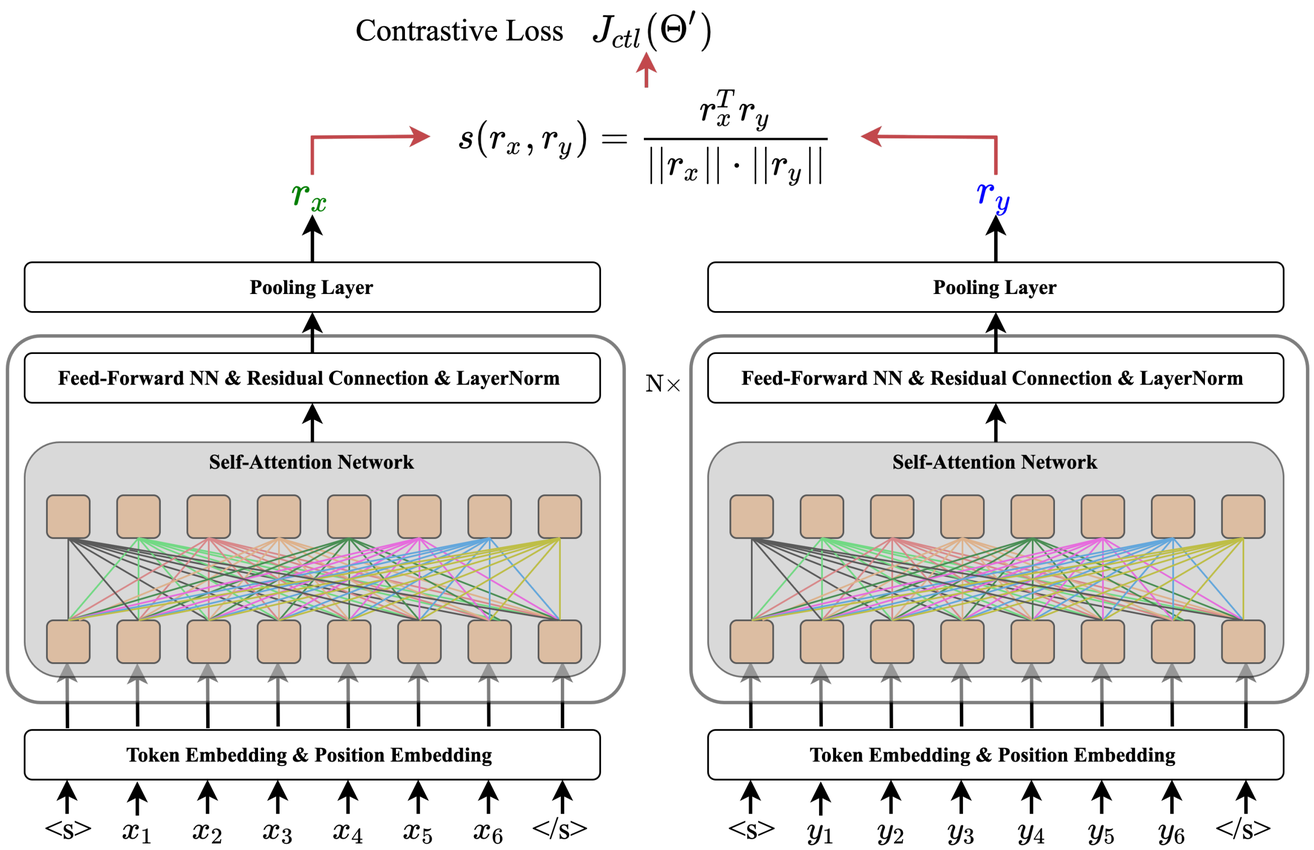


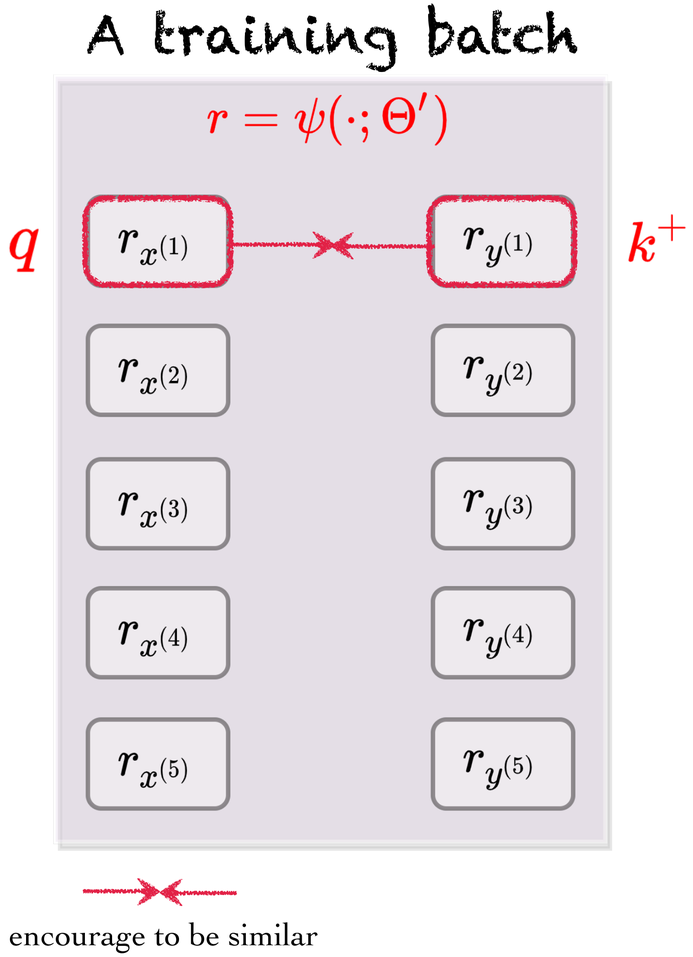


Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)

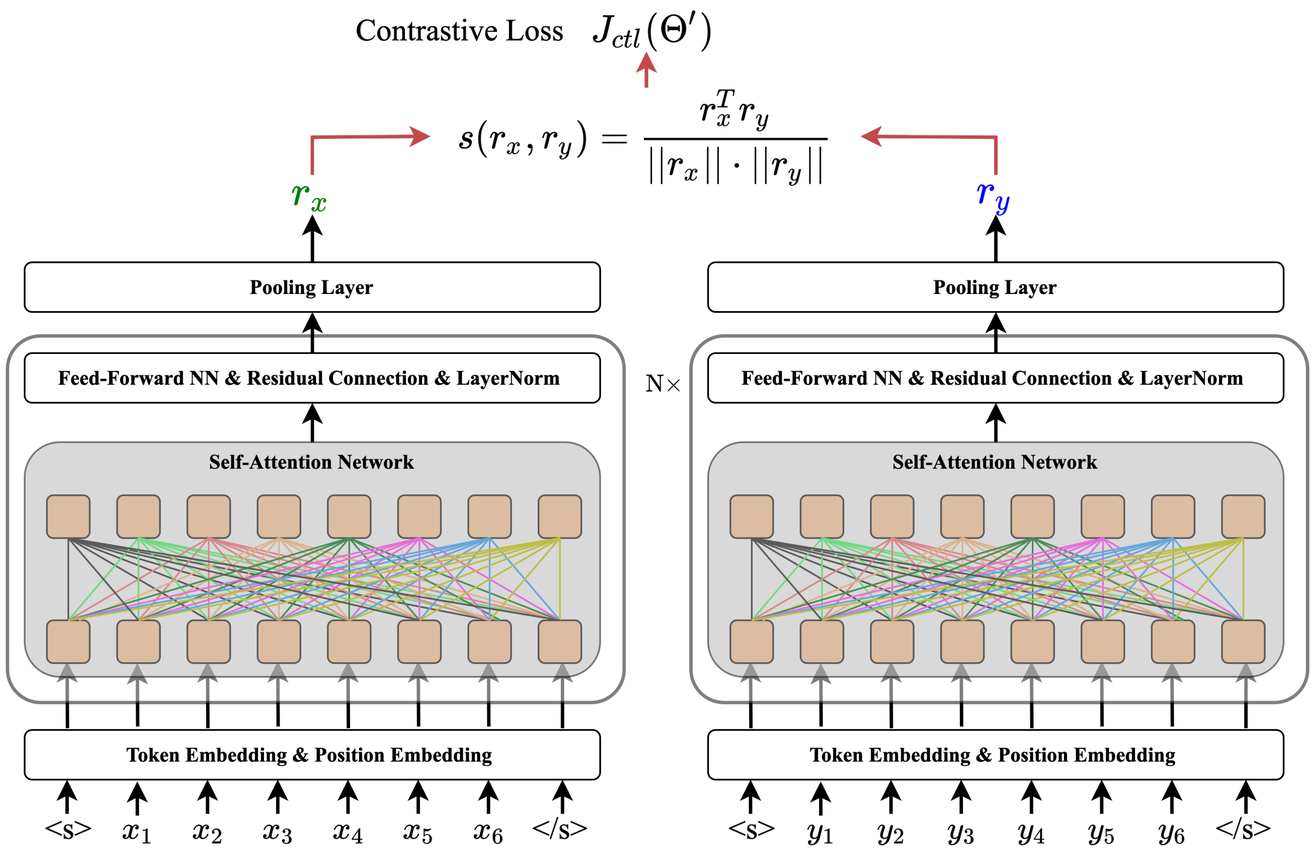


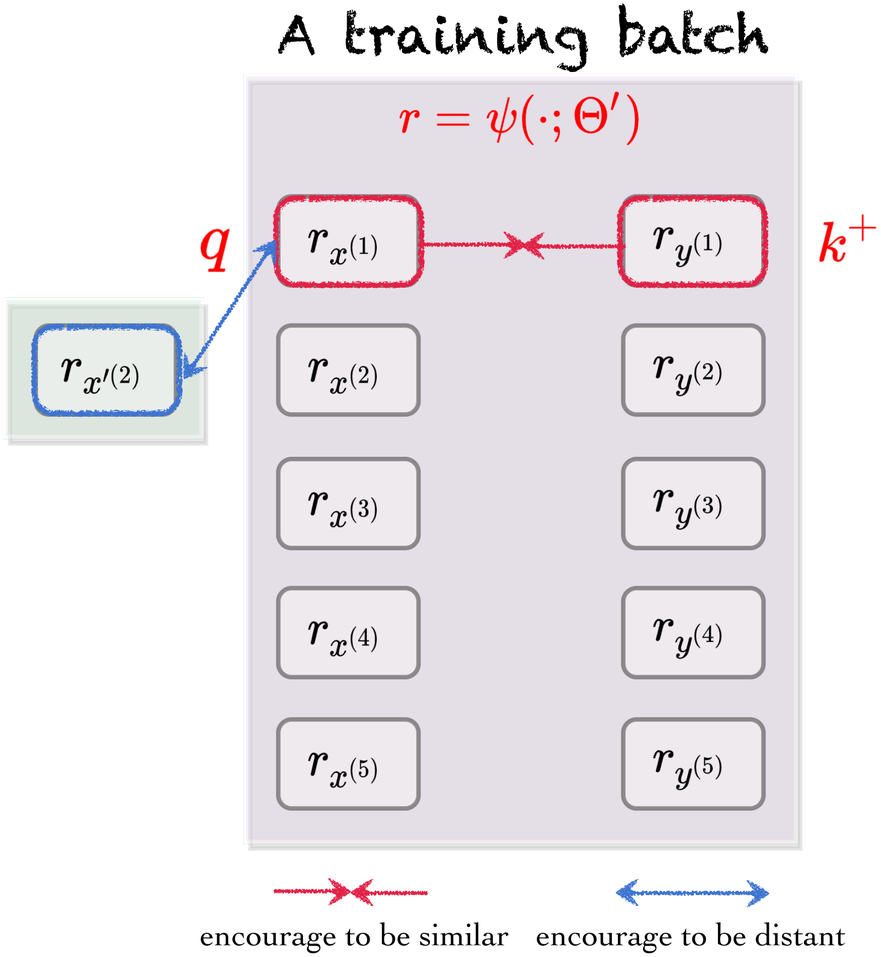


Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)

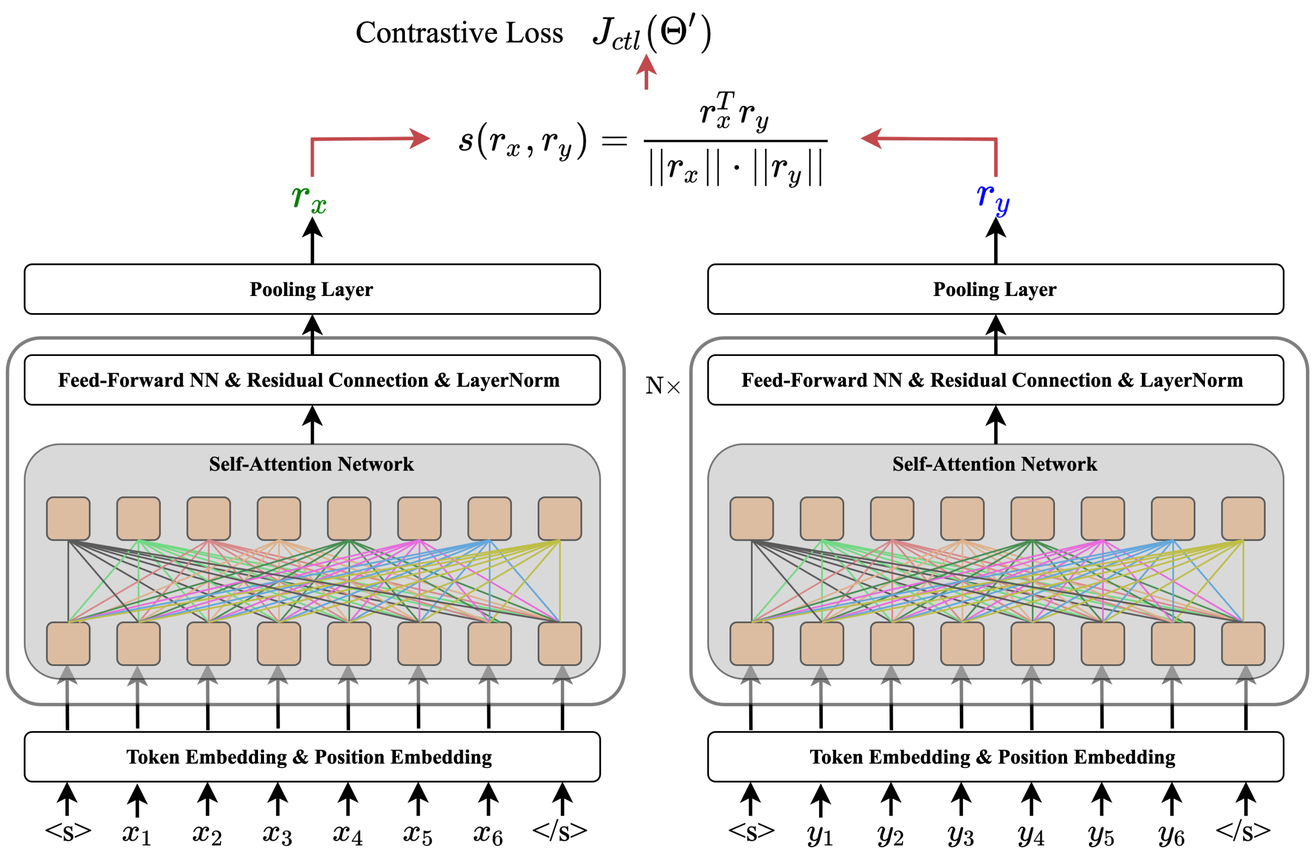


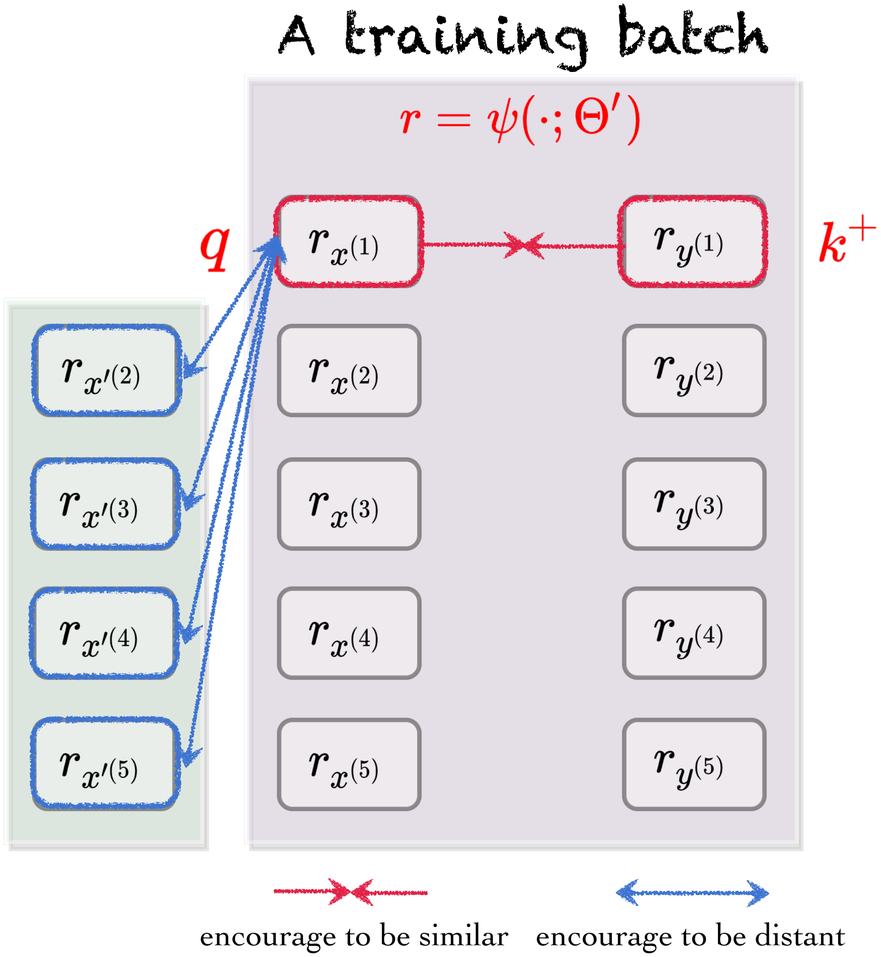


Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)

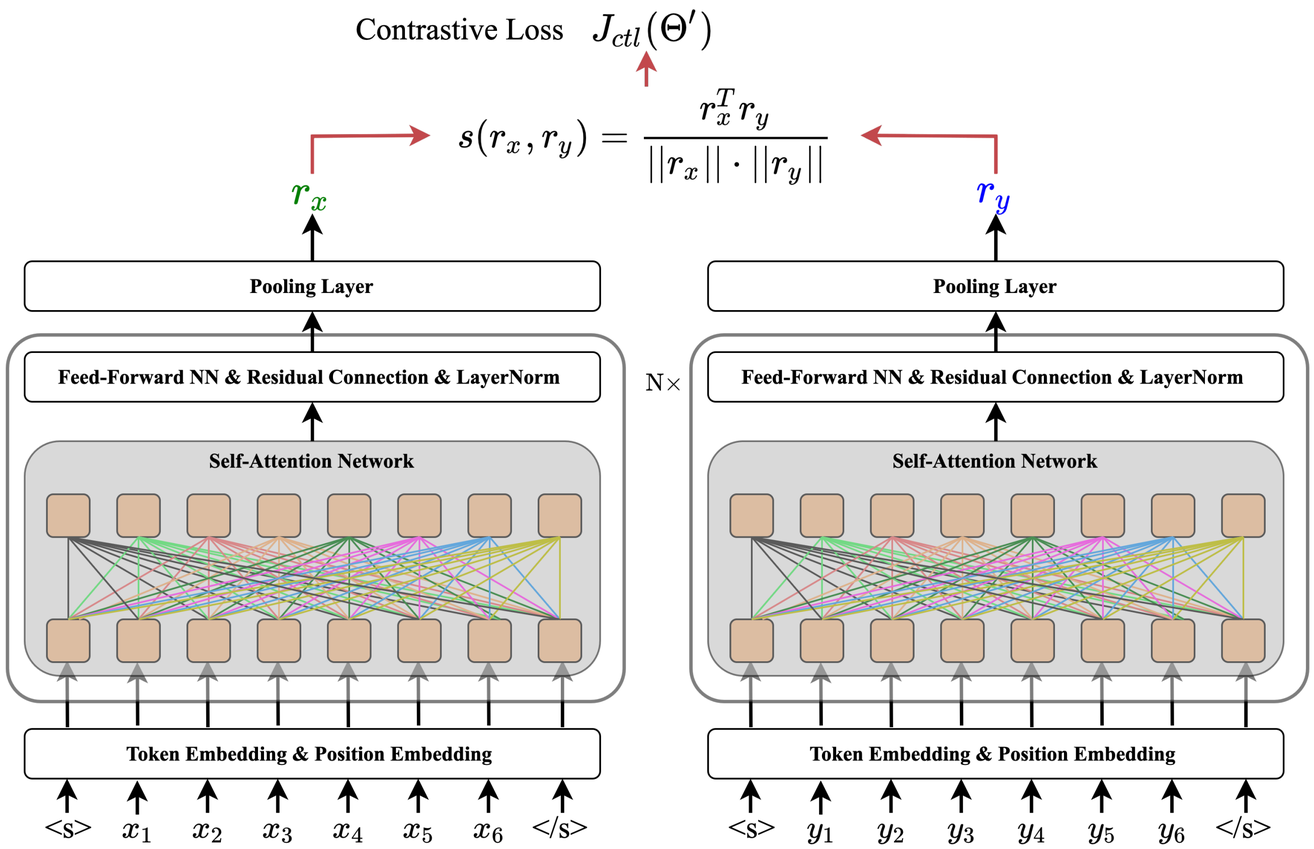


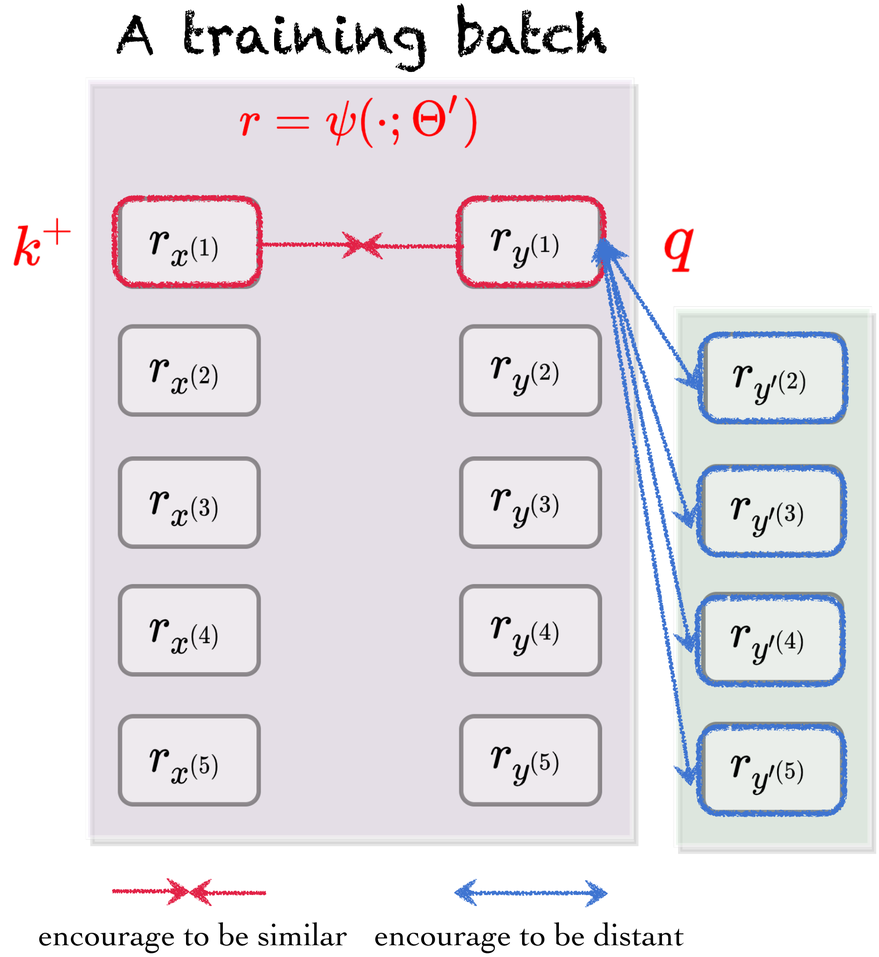


Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)

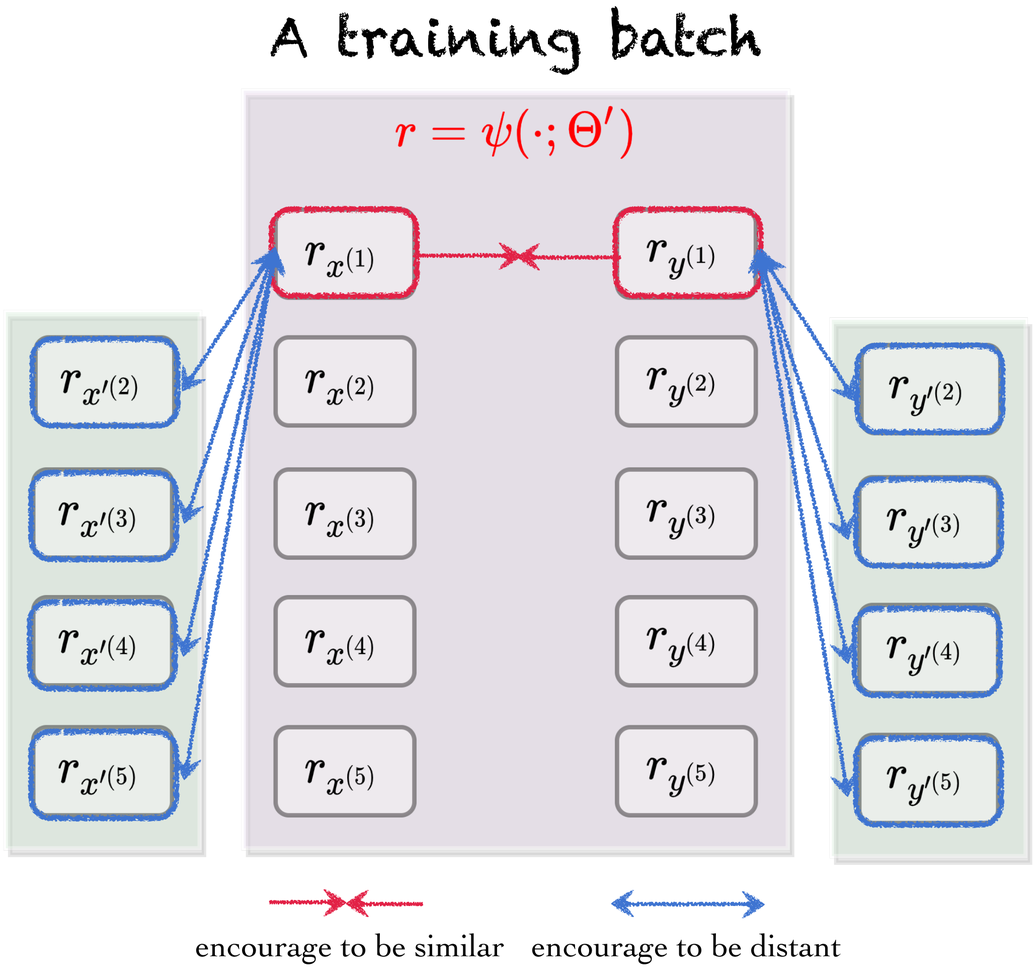




Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

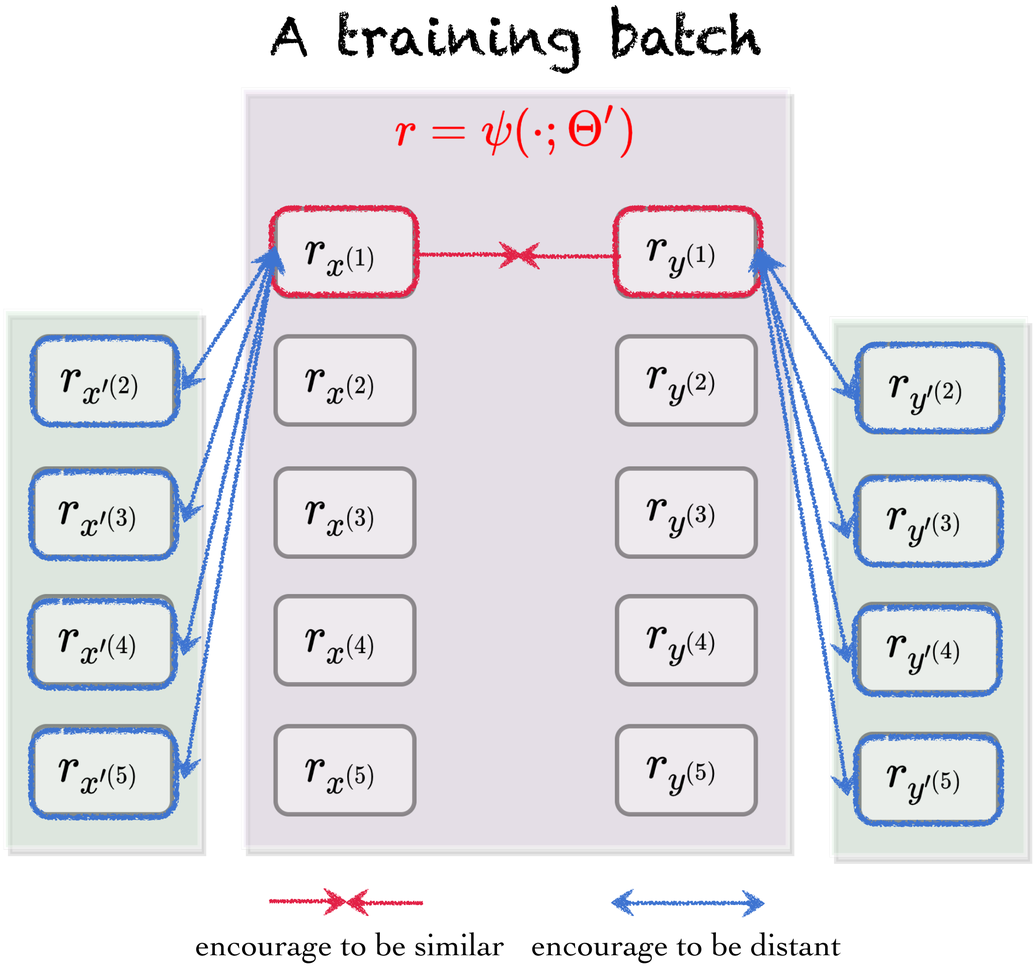
Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)



Tangential Contrastive Learning

Positive pairs: embeddings of each parallel two sentences are positives of each other

Negative pairs: (interpolated) embeddings of other sentences from the same mini-batch (in-batch negatives)



Parallel sentence pairs

$$J_{ctl}(\Theta') = \mathbb{E}_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\sim\mathcal{B}}\left(\log \frac{e^{s\left(r_{x^{(i)}}, r_{y^{(i)}}\right)}}{e^{s\left(r_{x^{(i)}}, r_{y^{(i)}}\right)} + \xi}\right)$$

$$\xi = \sum_{j\&j\neq i}^{|\mathcal{B}|}\left(e^{s\left(r_{y^{(i)}}, r_{y'^{(j)}}\right)} + e^{s\left(r_{x^{(i)}}, r_{x'^{(j)}}\right)}\right)$$

Interpolated in-batch negatives

Sampling with Mixed Gaussian Recurrent Chain

Our scenario is that each training instance always accompanies with a cluster of vectors sampled from its continuous vicinity $\nu(r_x, r_y)$, to enable the NMT model to generalize to more.

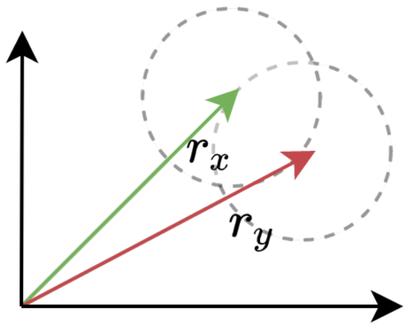
Challenges: the vicinity of each instance follows an unknown probability distribution.

Sampling with Mixed Gaussian Recurrent Chain

Our scenario is that each training instance always accompanies with a cluster of vectors sampled from its continuous vicinity $\nu(r_x, r_y)$, to enable the NMT model to generalize to more.

Challenges: the vicinity of each instance follows an unknown probability distribution.

A possible solution with vector transform

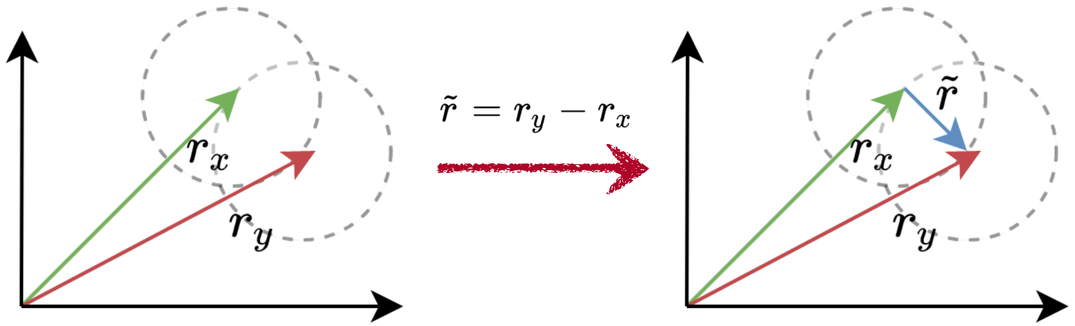


Sampling with Mixed Gaussian Recurrent Chain

Our scenario is that each training instance always accompanies with a cluster of vectors sampled from its continuous vicinity $\nu(r_x, r_y)$, to enable the NMT model to generalize to more.

Challenges: the vicinity of each instance follows an unknown probability distribution.

A possible solution with vector transform



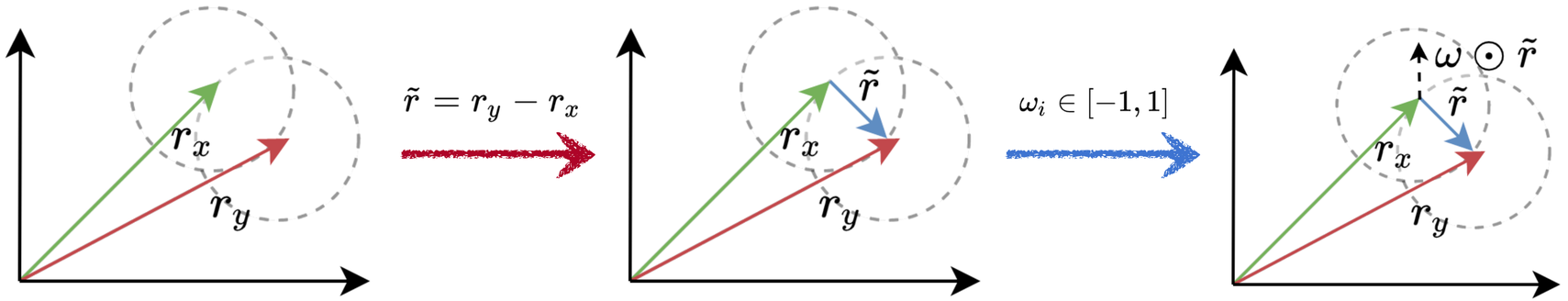
Compute the bias vector between the embeddings of two parallel sentences

Sampling with Mixed Gaussian Recurrent Chain

Our scenario is that each training instance always accompanies with a cluster of vectors sampled from its continuous vicinity $\nu(r_x, r_y)$, to enable the NMT model to generalize to more.

Challenges: the vicinity of each instance follows an unknown probability distribution.

A possible solution with vector transform



Compute the bias vector between the embeddings of two parallel sentences

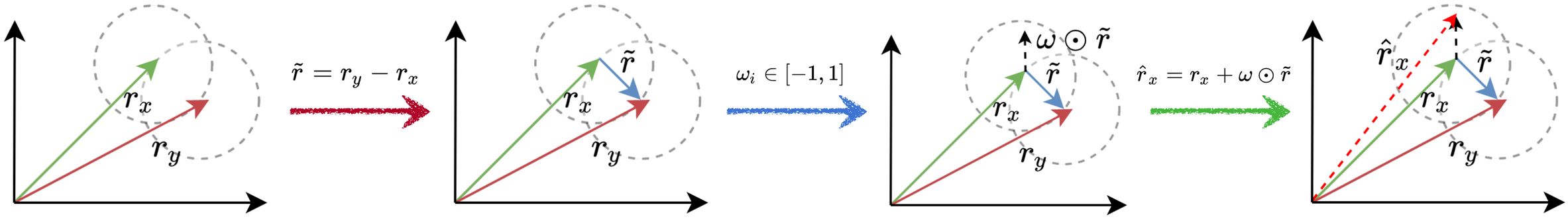
Transform the norm or the direction of the bias vector

Sampling with Mixed Gaussian Recurrent Chain

Our scenario is that each training instance always accompanies with a cluster of vectors sampled from its continuous vicinity $\nu(r_x, r_y)$, to enable the NMT model to generalize to more.

Challenges: the vicinity of each instance follows an unknown probability distribution.

A possible solution with vector transform



Compute the bias vector between the embeddings of two parallel sentences

Transform the norm or the direction of the bias vector

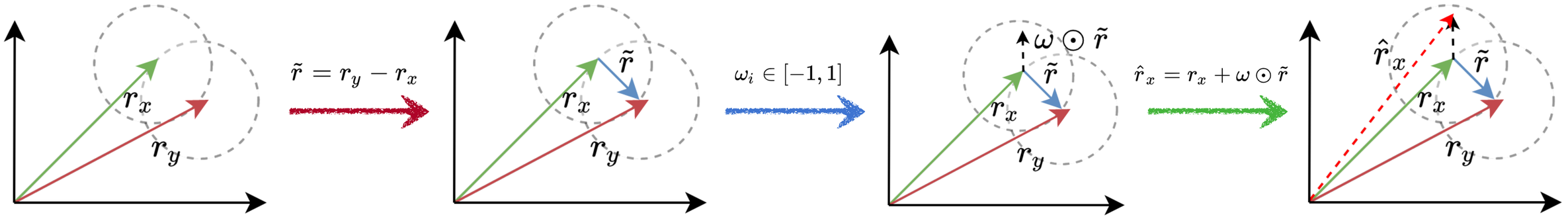
Newly construct a sample as the augmented instance

Sampling with Mixed Gaussian Recurrent Chain

Our scenario is that each training instance always accompanies with a cluster of vectors sampled from its continuous vicinity $\nu(r_x, r_y)$, to enable the NMT model to generalize to more.

Challenges: the vicinity of each instance follows an unknown probability distribution.

A possible solution with vector transform



As a consequence, the goal of the sampling strategy turns into find a set of scale vectors:

$$\omega \in \{\omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(K)}\}$$

Sampling with Mixed Gaussian Recurrent Chain

MGRC Sampling Algorithm

Input: The representations of the training instance $(\mathbf{x}, \mathbf{y})$, i.e. r_x and r_y.

Output: A set of augmented samples $\mathcal{R} = \{\hat{r}^{(1)}, \hat{r}^{(2)}, ..., \hat{r}^{(K)}\}$

Normalizing the importance of each element in $\tilde{r} = r_y - r_x$: $\mathcal{W}_r = \frac{|\tilde{r}| - min(|\tilde{r}|)}{max(|\tilde{r}|) - min(|\tilde{r}|)}$

Set $k = 1$, $\omega^{(1)} \sim \mathcal{N}(\mathbf{0}, \text{diag}(\mathcal{W}_r^2))$, $\hat{r}^{(1)} = r + \omega^{(1)} \odot (r_y - r_x)$

Initialize the set of samples as $\mathcal{R} = \{\hat{r}^{(1)}\}$.

while $k \leq (K - 1)$ **do**

  $k \leftarrow k + 1$

  Calculate the current scale vector: $\omega^{(k)} \sim p(\omega|\omega^{(1)}, \omega^{(2)}, ..., \omega^{(k-1)}$ according to Eq. (6).

  Calculate the current sample: $\hat{r}^{(k)} = r + \omega^{(k)} \odot (r_y - r_x)$.

  $\mathcal{R} \leftarrow \mathcal{R} \bigcup \{\hat{r}^{(k)}\}$.

end while

Sampling with Mixed Gaussian Recurrent Chain

MGRC Sampling Algorithm

Input: The representations of the training instance $(\mathbf{x}, \mathbf{y})$, i.e. r_x and r_y.

Output: A set of augmented samples $\mathcal{R} = \{\hat{r}^{(1)}, \hat{r}^{(2)}, ..., \hat{r}^{(K)}\}$

1: Normalizing the importance of each element in $\tilde{r} = r_y - r_x$: $\mathcal{W}_r = \frac{|\tilde{r}| - min(|\tilde{r}|)}{max(|\tilde{r}|) - min(|\tilde{r}|)}$

2: Set $k = 1$, $\omega^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\mathcal{W}_r^2))$, $\hat{r}^{(1)} = r + \omega^{(1)} \odot (r_y - r_x)$

3: Initialize the set of samples as $\mathcal{R} = \{\hat{r}^{(1)}\}$.

4: **while** $k \leq (K-1)$ **do**

5: $k \leftarrow k + 1$

6: Calculate the current scale vector: $\omega^{(k)} \sim p(\omega|\omega^{(1)}, \omega^{(2)}, ..., \omega^{(k-1)}$ according to Eq. (6).

7: Calculate the current sample: $\hat{r}^{(k)} = r + \omega^{(k)} \odot (r_y - r_x)$.

8: $\mathcal{R} \leftarrow \mathcal{R} \bigcup \{\hat{r}^{(k)}\}$.

9: **end while**

Limits the range of sampling to a subspace of the vicinity, and rejects to conduct sampling from the uninformative dimensions

$$p = \eta\mathcal{N}\left(\mathbf{0}, \mathrm{diag}(\mathcal{W}_r^2)\right) + (1.0 - \eta)\mathcal{N}\left(\frac{1}{k-1}\sum_{i=1}^{k-1}\omega^{(i)}, \mathbf{1}\right)$$

Describes the fact that the diversity of each training instance is not infinite

Training Strategy

① Tangential contrastive learning with adaptive negatives to optimize the semantic encoder:

$$\Theta'^{*} = \text{argmax}_{\Theta'} J_{ctl}(\Theta')$$

② Maximum likelihood estimation with vicinities to optimize the NMT model:

$$\Theta^{*} = \text{argmax}_{\Theta} J_{mle}(\Theta)$$

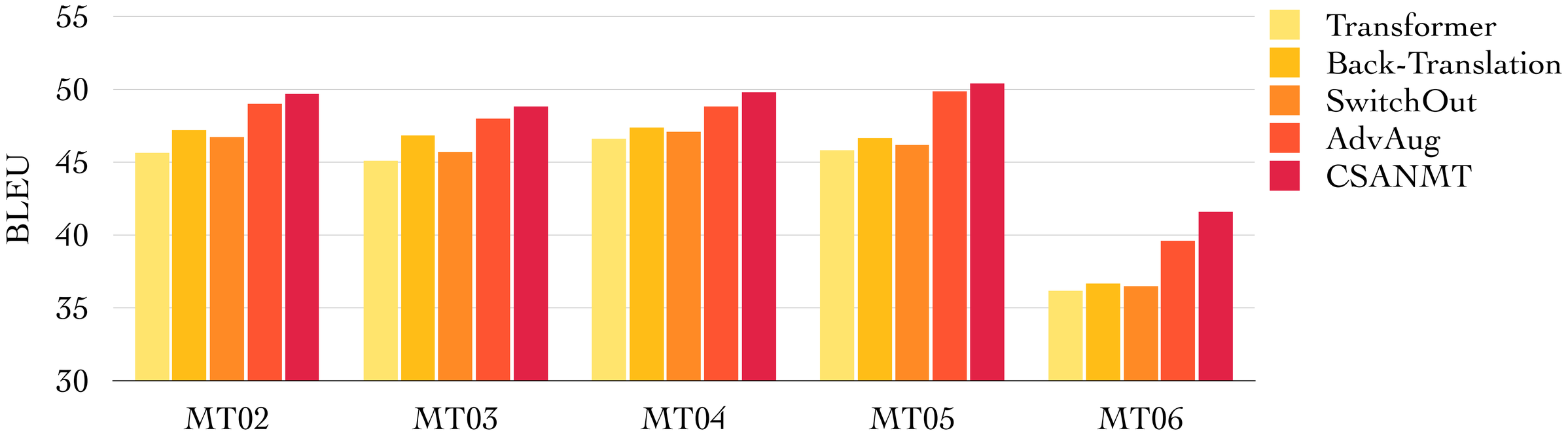
③ Jointly optimization while use a small learning rate to fine-tune the semantic encoder

$$\Theta = \Theta - \alpha_{mle} \cdot \nabla_{\Theta} J_{mle}(\Theta) \qquad \Theta' = \Theta' - \alpha_{ctl} \cdot \nabla_{\Theta'} J_{ctl}(\Theta')$$

a very small learning rate, e.g. 1e-5

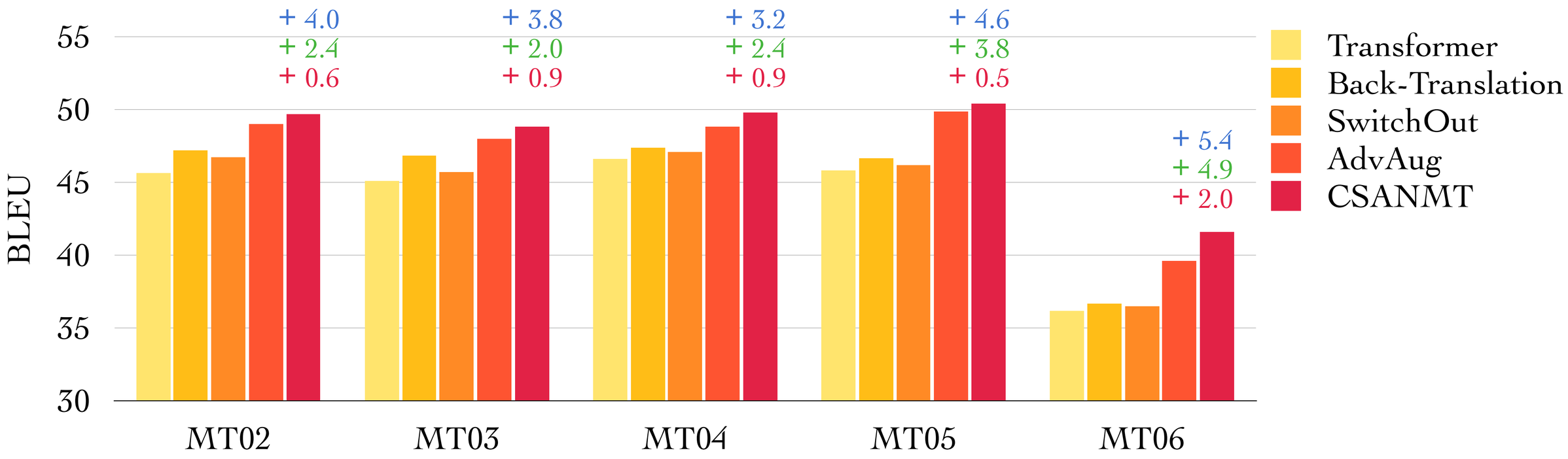
CSANMT: Main Results

NIST Chinese-to-English, LDC corpus (1.25M) for training, MT06 (valid), MT02-05&08 (test)



CsaNMT: Main Results

NIST Chinese-to-English, LDC corpus (1.25M) for training, MT06 (valid), MT02-05&08 (test)



4.2 BLEU higher than Transformer

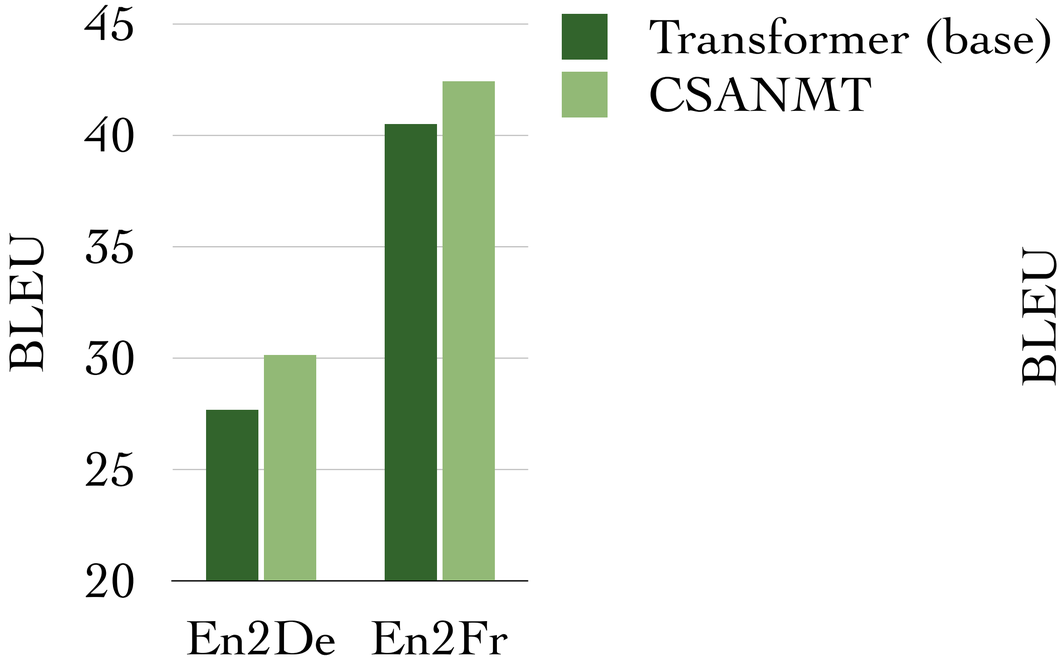
3.1 BLEU higher than Back-Translation

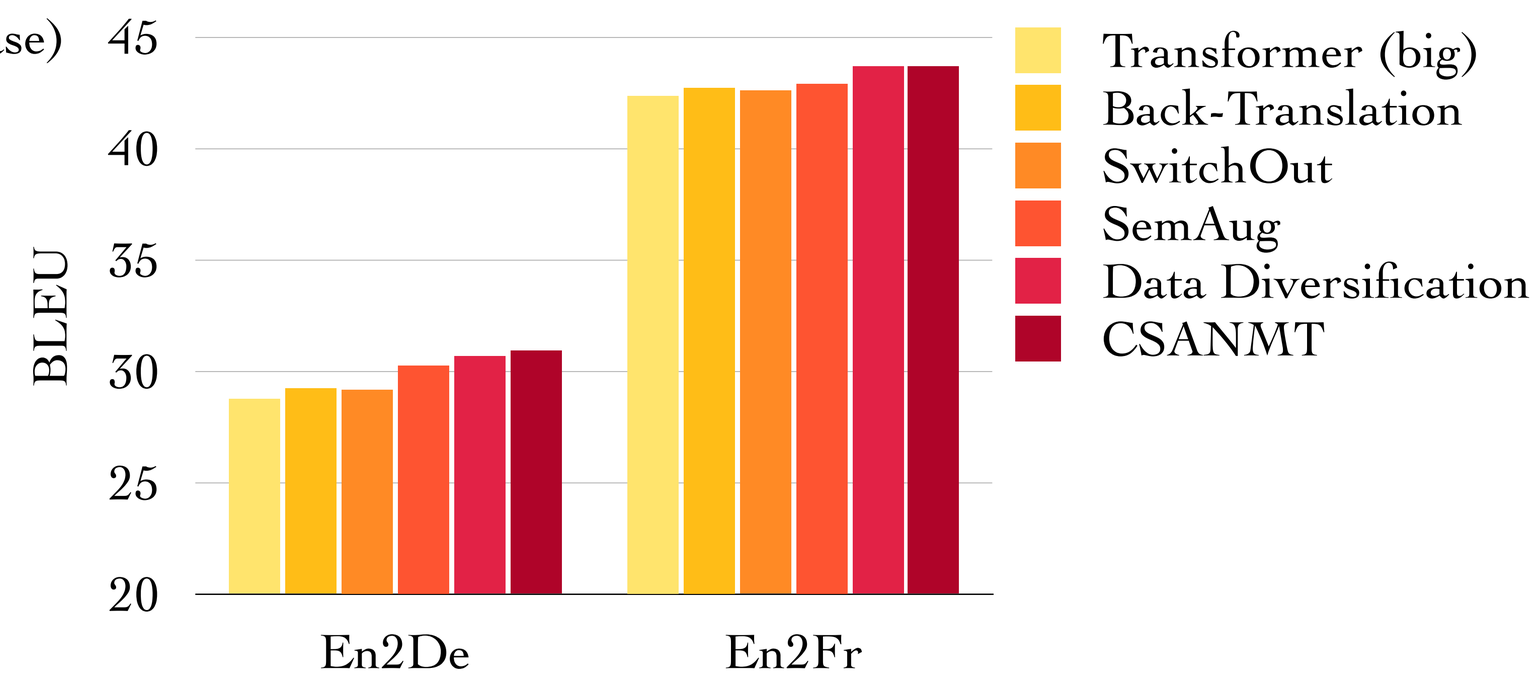
1.0 BLEU higher than AdvAug

CSANMT: Main Results

English-German, WMT'14 corpus (4.5M) for training, newstest2013 (valid), newstest2014 (test)

English-French, WMT'14 corpus (36M) for training, newstest2012 & 2013 (valid), newstest2014 (test)

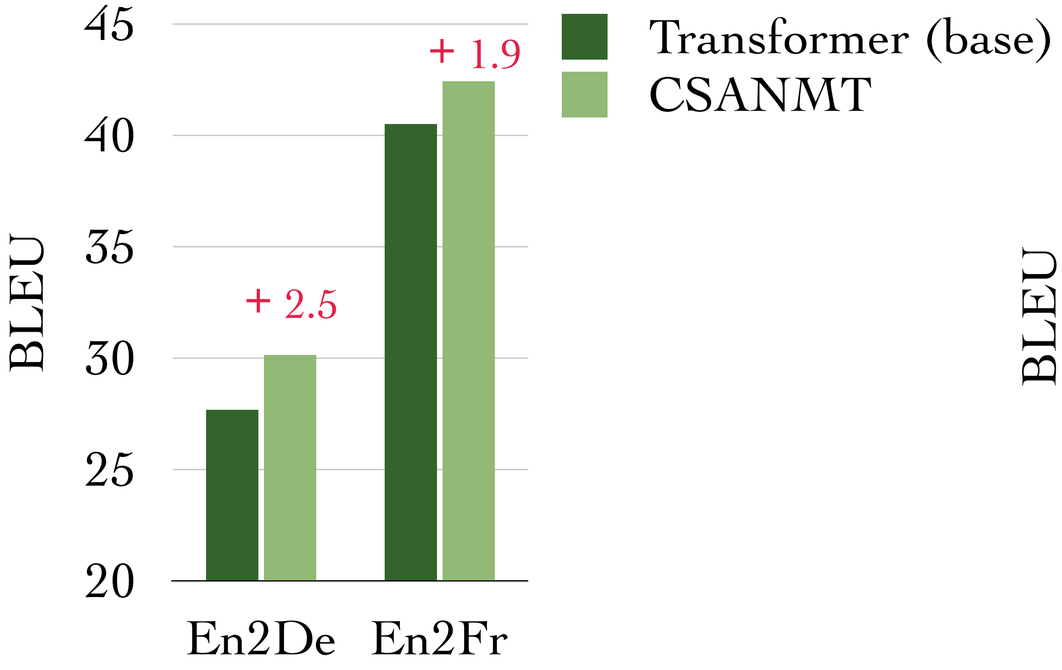


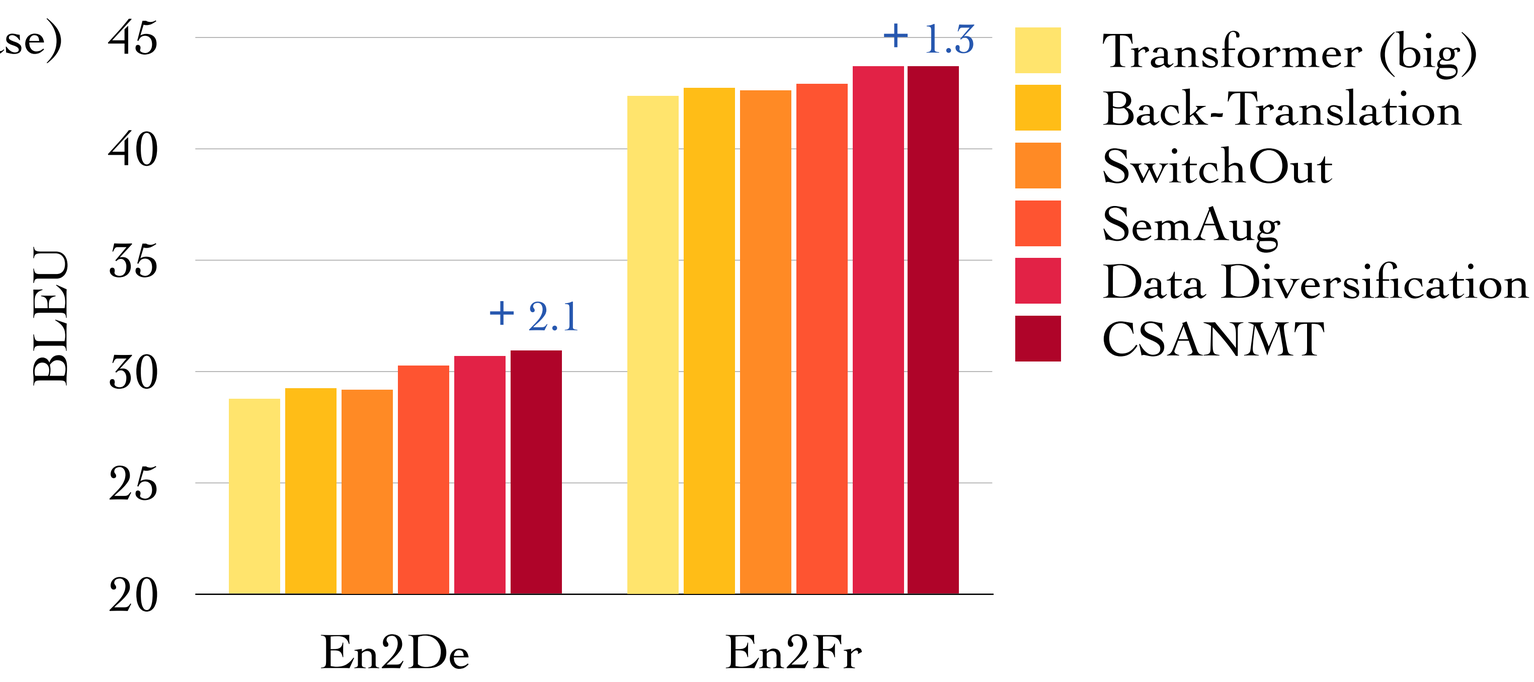


CSANMT: Main Results

English-German, WMT'14 corpus (4.5M) for training, newstest2013 (valid), newstest2014 (test)

English-French, WMT'14 corpus (36M) for training, newstest2012 & 2013 (valid), newstest2014 (test)





Significantly improves the vanilla Transformer, and performs comparably to the previous SOTA.

See more results in the paper

Why Does this Work?

Learning objective of semantic encoder and sampling algorithm

Tangential CTL

$$J_{ctl}(\Theta') = \mathbb{E}_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\sim\mathcal{B}}\left(\log\frac{e^{s\left(r_{x^{(i)}},r_{y^{(i)}}\right)}}{e^{s\left(r_{x^{(i)}},r_{y^{(i)}}\right)}+\xi}\right)$$

MGRC

$$p = \eta\mathcal{N}\left(\mathbf{0},\mathrm{diag}(\mathcal{W}_r^2)\right) + (1.0-\eta)\mathcal{N}\left(\frac{1}{k-1}\sum_{i=1}^{k-1}\omega^{(i)},\mathbf{1}\right)$$

Why Does this Work?

Learning objective of semantic encoder and sampling algorithm

Tangential CTL

$$J_{ctl}(\Theta') = \mathbb{E}_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\sim\mathcal{B}}\left(\log\frac{e^{s\left(r_{x^{(i)}},r_{y^{(i)}}\right)}}{e^{s\left(r_{x^{(i)}},r_{y^{(i)}}\right)}+\xi}\right)$$

MGRC

$$p = \eta\mathcal{N}\left(\mathbf{0},\mathrm{diag}(\mathcal{W}_r^2)\right) + (1.0-\eta)\mathcal{N}\left(\frac{1}{k-1}\sum_{i=1}^{k-1}\omega^{(i)},\mathbf{1}\right)$$

Compare it to

Tangential CTL

$$J_{ctl}(\Theta') = \mathbb{E}_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\sim\mathcal{B}}\left(\log\frac{e^{s\left(r_{x^{(i)}},r_{y^{(i)}}\right)}}{e^{s\left(r_{x^{(i)}},r_{y^{(i)}}\right)}+\xi}\right)$$

MGRC w/o recurrent chain

$$p = \eta\mathcal{N}\left(\mathbf{0},\mathrm{diag}(\mathcal{W}_r^2)\right) + (1.0-\eta)\mathcal{N}\left(\frac{1}{k-1}\sum_{i=1}^{k-1}\omega^{(i)},\mathbf{1}\right)$$

Tangential CTL

$$J_{ctl}(\Theta') = \mathbb{E}_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\sim\mathcal{B}}\left(\log\frac{e^{s\left(r_{x^{(i)}},r_{y^{(i)}}\right)}}{e^{s\left(r_{x^{(i)}},r_{y^{(i)}}\right)}+\xi}\right)$$

Replace Gaussian with uniform dist.

$$\omega^{(k)} \sim \eta\mathcal{U}\left(-\mathcal{W}_r,\mathcal{W}_r\right) + (1.0-\eta)\mathcal{U}\left(\bar{\mathbf{a}}-\mathbf{1},\mathbf{1}-\bar{\mathbf{a}}\right)$$

Variational Inference

$$\mathbb{E}_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\sim\mathcal{B}}\left(-KL\left(p(r_{x^{(i)}};\mu,\sigma^2)\parallel q(r_{x^{(i)}},r_{y^{(i)}};\mu',\sigma'^2)\right)\right)$$

Reparameterization

$$\hat{r}_x = \mu + \epsilon\odot\sigma \qquad \epsilon\sim\mathcal{N}(0,1)$$

Cosine similarity

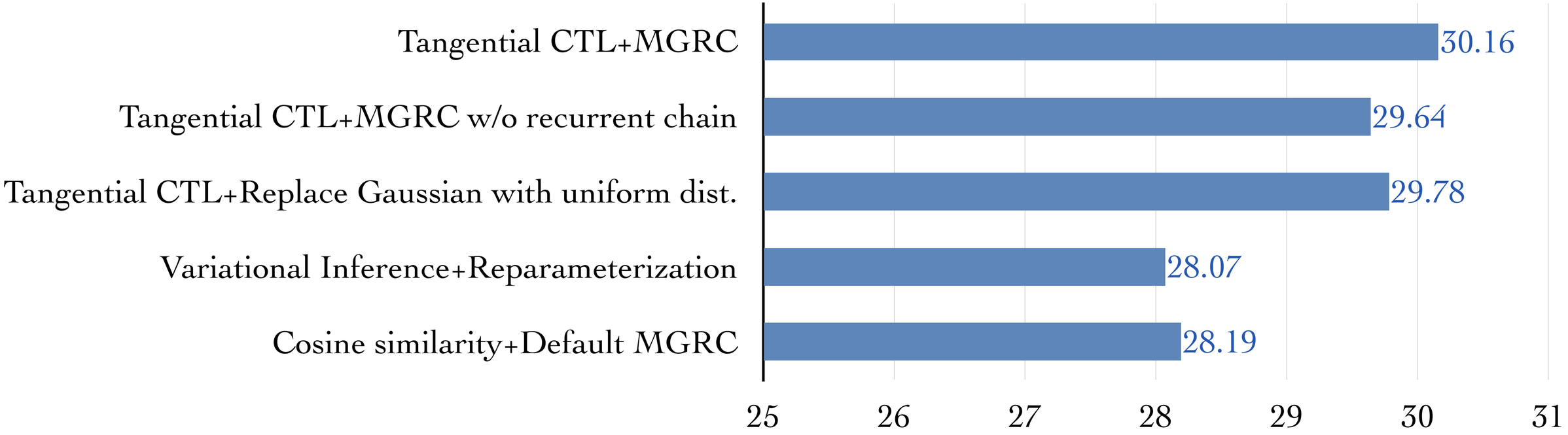
$$\mathbb{E}_{(\mathbf{x}^{(i)},\mathbf{y}^{(i)})\sim\mathcal{B}}\left(\frac{r_{x^{(i)}}^T r_{y^{(i)}}}{\|r_{x^{(i)}}\|\cdot\|r_{y^{(i)}}\|}\right)$$

Default MGRC

$$p = \eta\mathcal{N}\left(\mathbf{0},\mathrm{diag}(\mathcal{W}_r^2)\right) + (1.0-\eta)\mathcal{N}\left(\frac{1}{k-1}\sum_{i=1}^{k-1}\omega^{(i)},\mathbf{1}\right)$$

Why Does this Work?

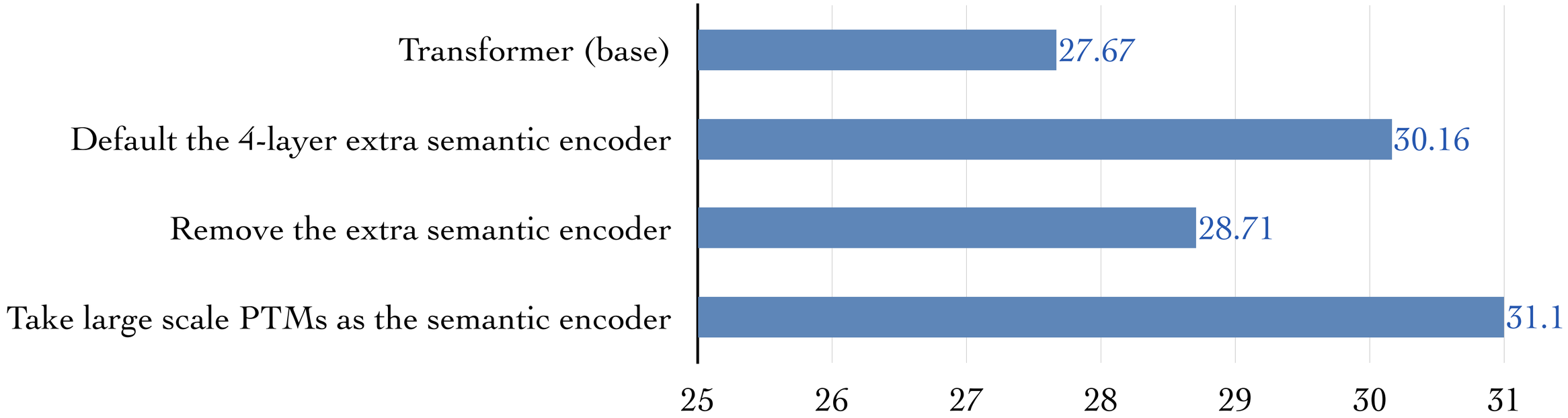
Learning objective of semantic encoder and sampling algorithm



- Tangential CTL ≫ Cosine similarity / Variational Inference (!)
- Tangential CTL + MGRC ≫ Variational Inference + Reparameterization (previous continuous generation)

Why Does this Work?

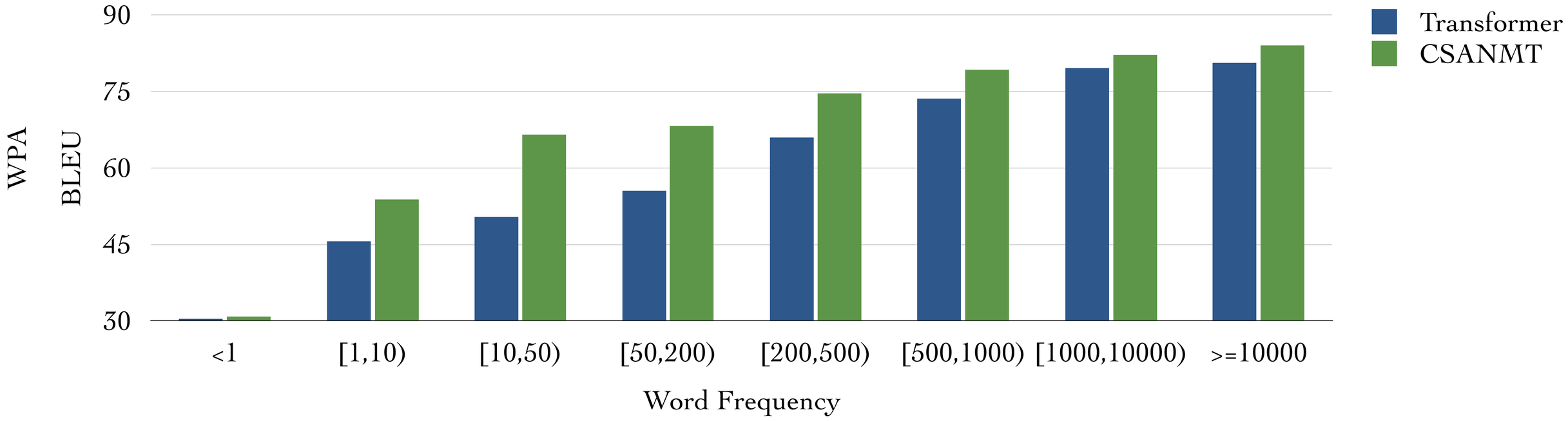
Effect of the semantic encoder variants



- The extra semantic encoder is necessary even it is small
- Take large scale pre-trained models (i.e., XLM-R) as the semantic encoder further improves ~1.0 BLEU

Why Does this Work?

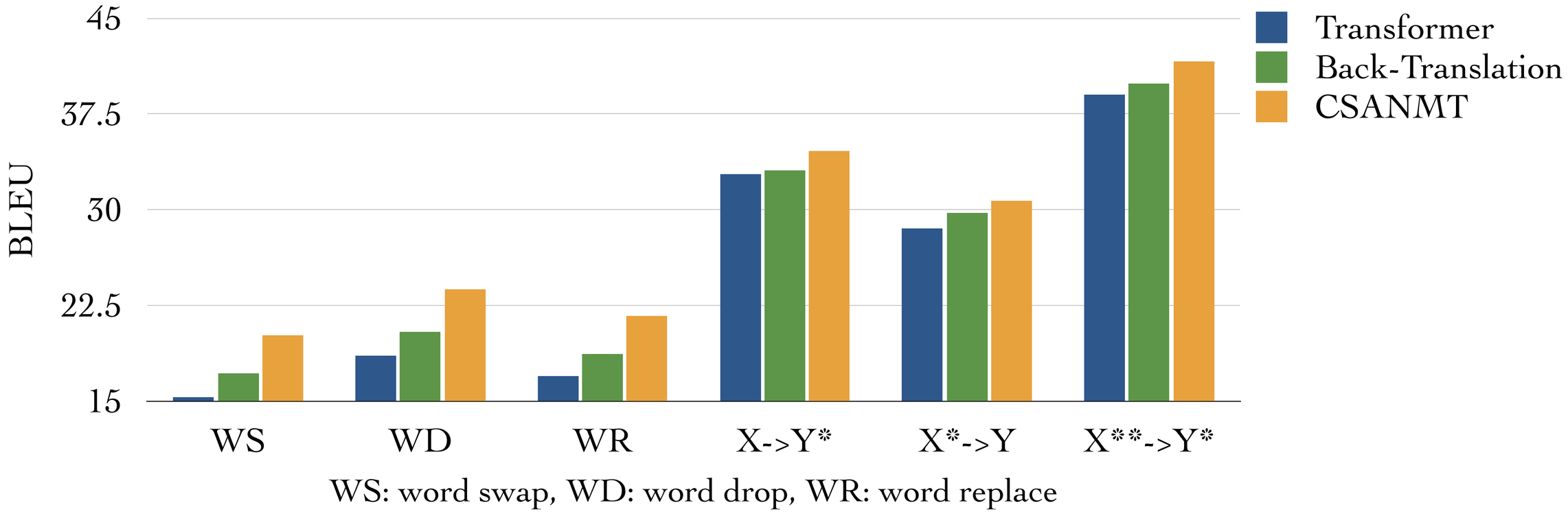
Word Prediction Accuracy (WPA)



- CSANMT generalizes to rare words better than the vanilla Transformer (gap of WPA up to 16%)

Why Does this Work?

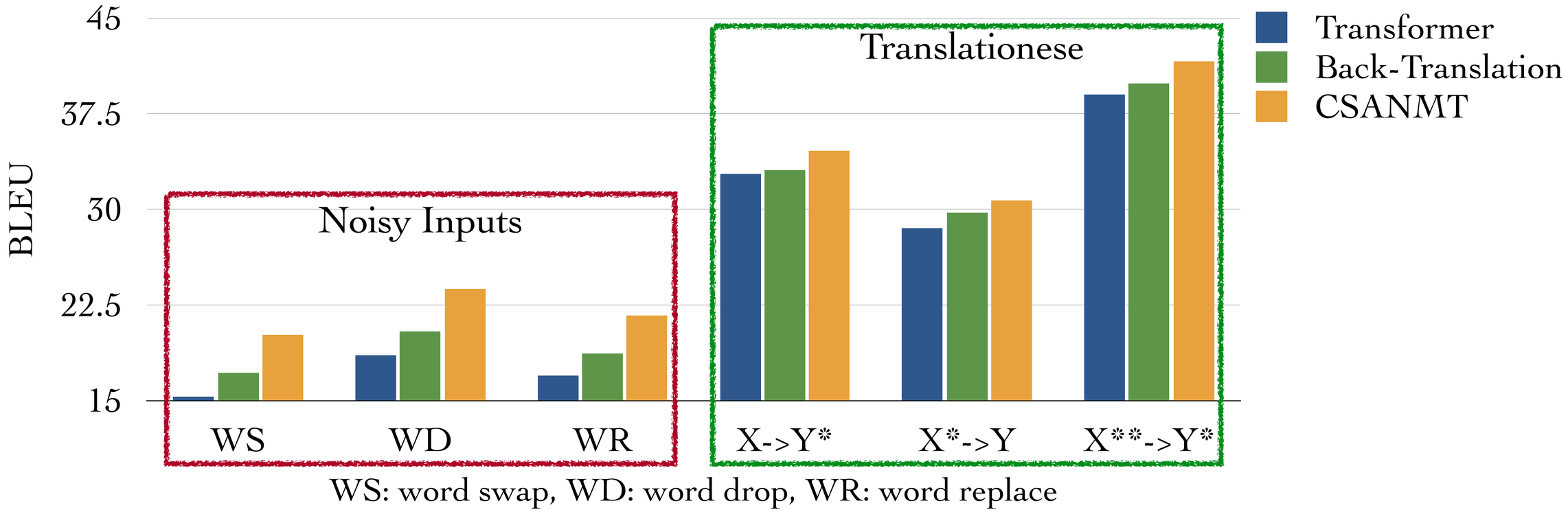
Robustness towards Noisy Inputs and Translationese



WS: word swap, WD: word drop, WR: word replace

Why Does this Work?

Robustness towards Noisy Inputs and Translationese



WS: word swap, WD: word drop, WR: word replace

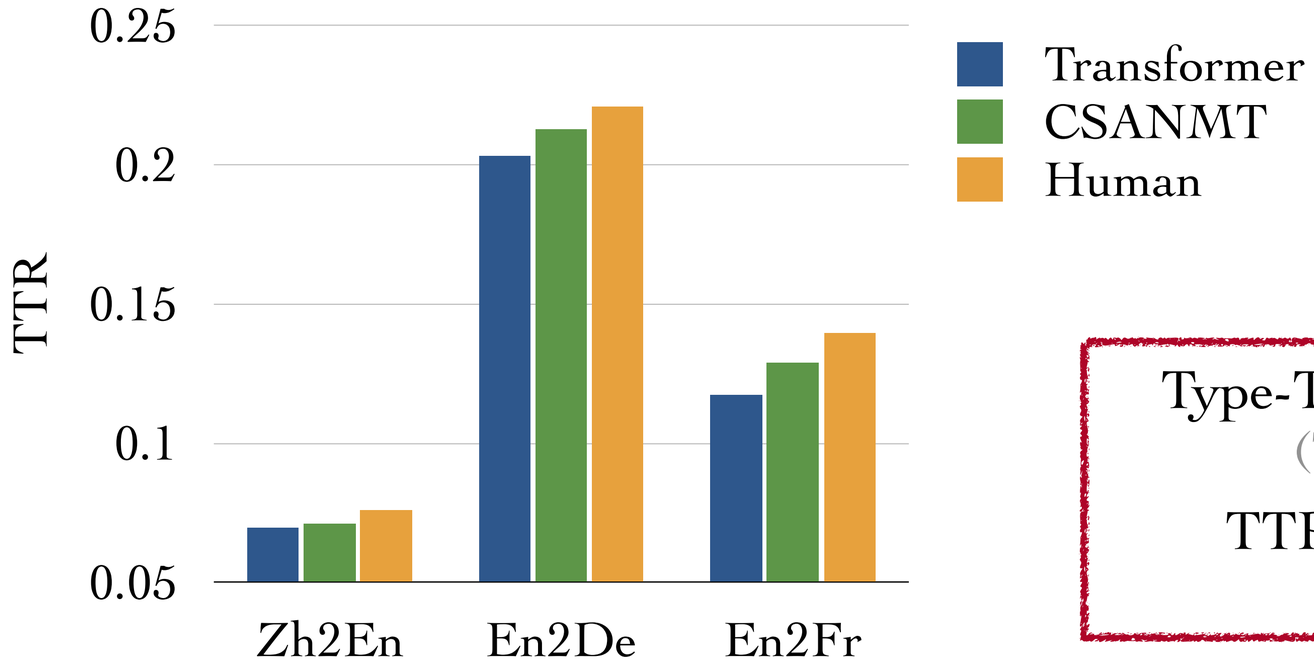
- Continuous semantic augmentation ≫ discrete data augmentation (i.e. back-translation)
- Continuous semantic augmentation benefits more (than discrete data augmentation) from natural parallel data to deal with the translationese inputs

Why Does this Work?

Lexical Diversity and Semantic Faithfulness

Why Does this Work?

Lexical Diversity and Semantic Faithfulness



Type-Token-Ratio (TTR)
(Templin, 1957)

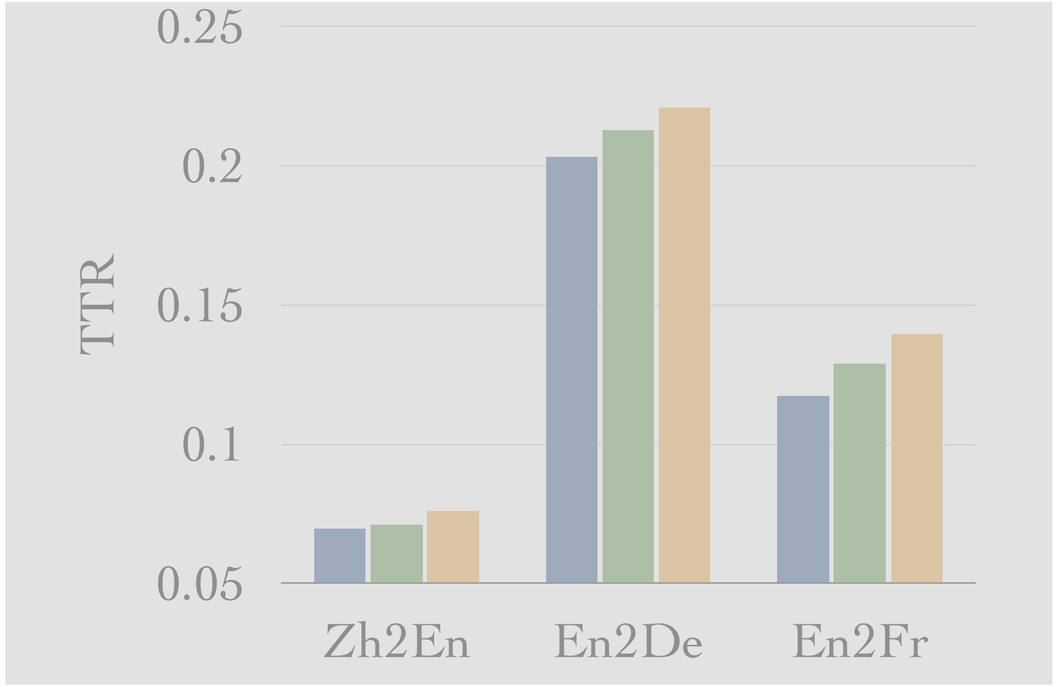
$$\text{TTR} = \frac{num.\ of\ types}{num.\ of\ tokens}$$

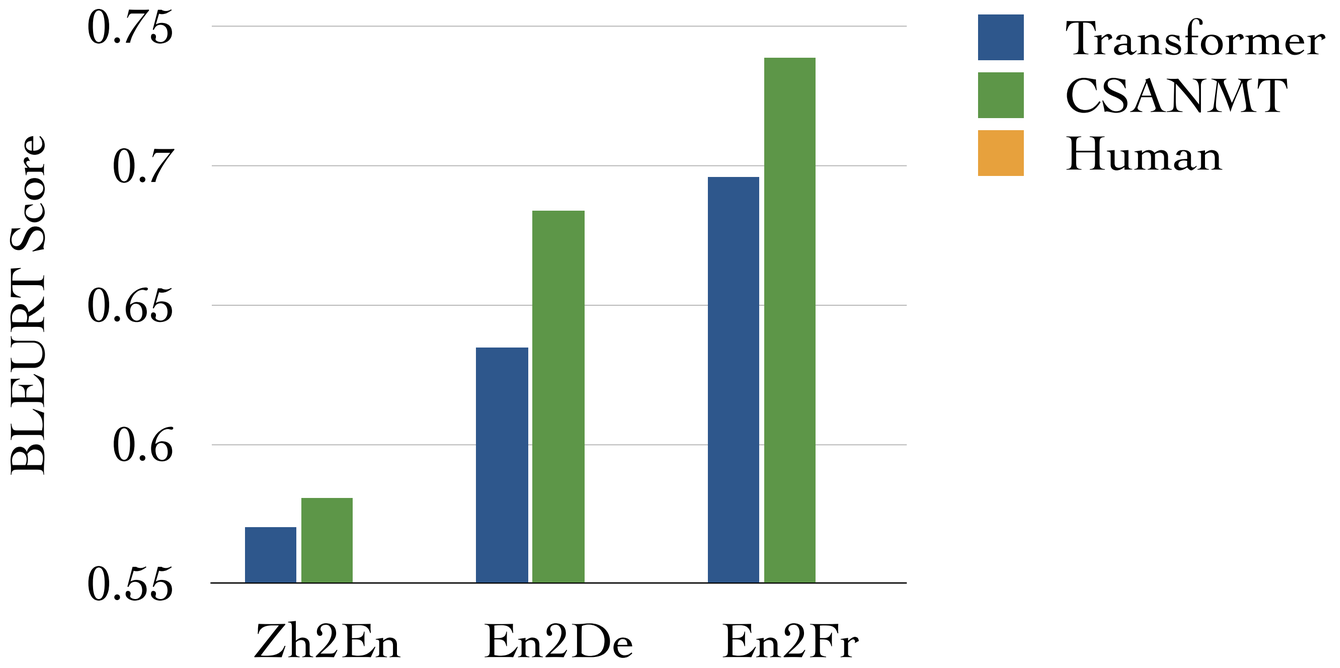
- Continuous semantic augmentation substantially bridges the gap of the lexical diversity between translations produced by human and machine.

Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.

Why Does this Work?

Lexical Diversity and Semantic Faithfulness

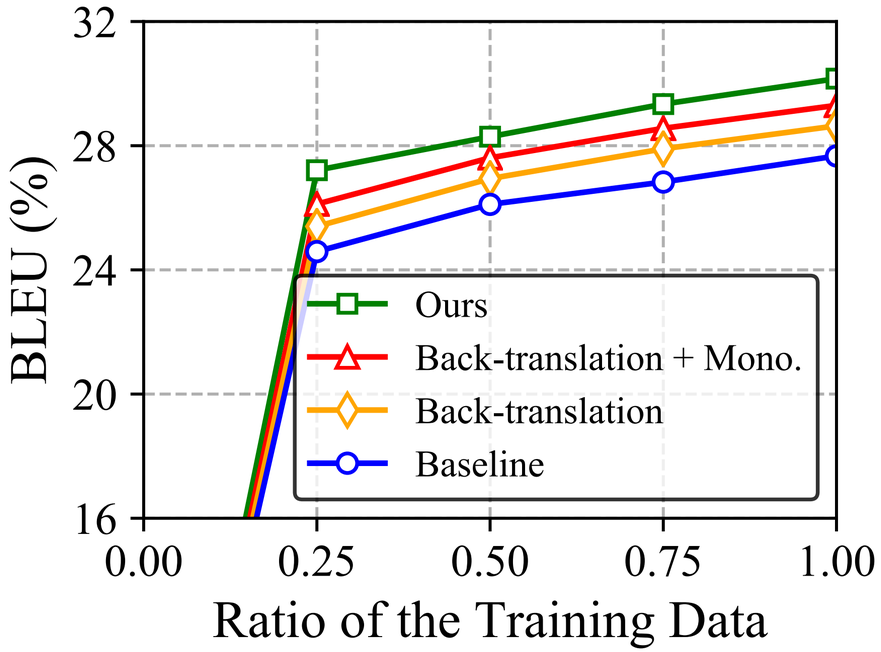




- Continuous semantic augmentation substantially bridges the gap of the lexical diversity between translations produced by human and machine.
- Continuous semantic augmentation shows a better capability on preserving the semantics of the generated translations than Transformer

Why Does this Work?

Data utilization and training efficiency

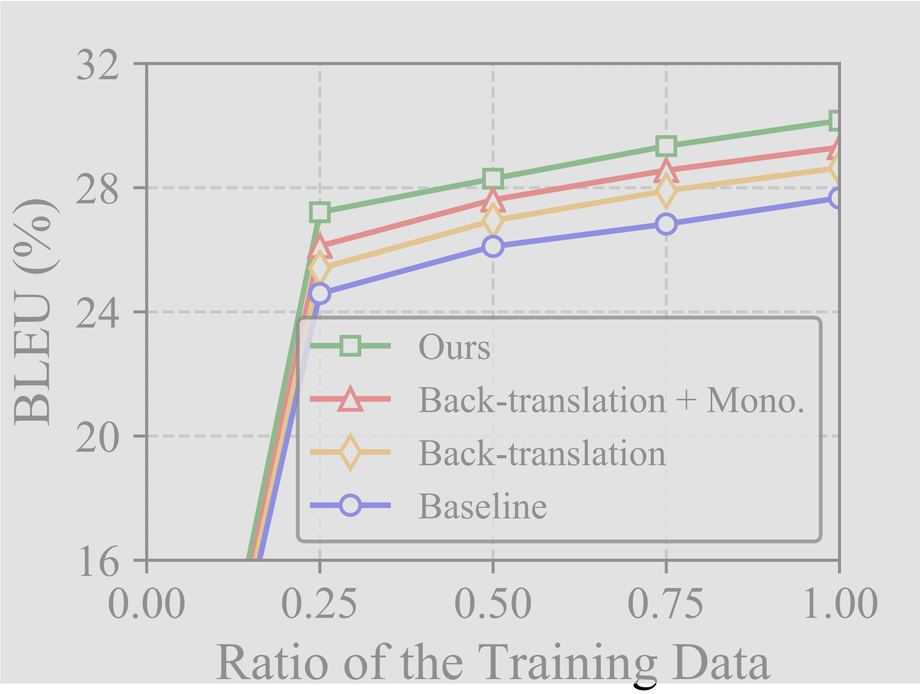


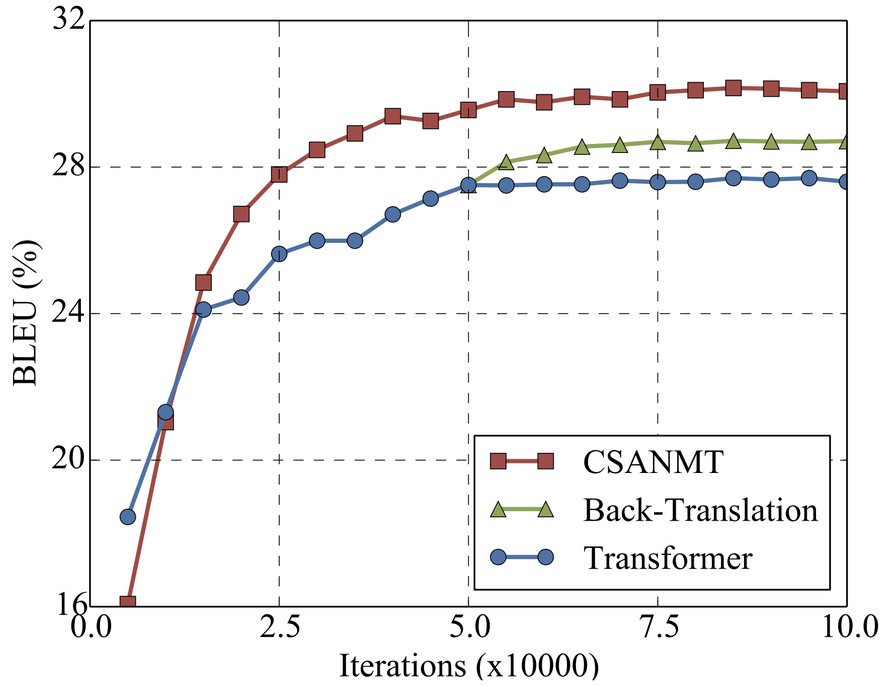
CSANMT achieves comparable performance with the baseline model with only 25% of training data, which indicates that our approach has great potential to achieve good results with very few data.

- Continuous semantic augmentation (without any extra monolingual data) consistently outperforms discrete data augmentation (even with extra 2M monolingual data) across different data ratios.

Why Does this Work?

Data utilization and training efficiency





We need 44% more training costs than the vanilla Transformer, for the longer time to make the CSANMT model converge

- Continuous semantic augmentation (without any extra monolingual data) consistently outperforms discrete data augmentation (even with extra 2M monolingual data) across different data ratios.
- Continuous semantic augmentation performs consistently better than both the vanilla Transformer and the back-translation method at each iteration.

Summary

CSANMT A novel data augmentation paradigm for NMT, which achieves state-of-the-art results

Summary

CSANMT A novel data augmentation paradigm for NMT, which achieves state-of-the-art results

- **Tangential Contrastive Learning**: pre-train a semantic encoder, which forms an adjacency semantic region (i.e. the *vicinity*, cover adequate variants of literal expression under the same meaning) for each training example.

Summary

CSANMT A novel data augmentation paradigm for NMT, which achieves state-of-the-art results

- **Tangential Contrastive Learning**: pre-train a semantic encoder, which forms an adjacency semantic region (i.e. the *vicinity*, cover adequate variants of literal expression under the same meaning) for each training example.
- **Mixed Gaussian Recurrent Chain**: produce a series of diverse training data for each observed instance from its vicinity, i.e. a semantically-preserved continuous sub-space.

Summary

CsaNMT A novel data augmentation paradigm for NMT, which achieves state-of-the-art results

- **Tangential Contrastive Learning**: pre-train a semantic encoder, which forms an adjacency semantic region (i.e. the *vicinity*, cover adequate variants of literal expression under the same meaning) for each training example.
- **Mixed Gaussian Recurrent Chain**: produce a series of diverse training data for each observed instance from its vicinity, i.e. a semantically-preserved continuous sub-space.

Next Step

Summary

CsaNMT A novel data augmentation paradigm for NMT, which achieves state-of-the-art results

- **Tangential Contrastive Learning**: pre-train a semantic encoder, which forms an adjacency semantic region (i.e. the *vicinity*, cover adequate variants of literal expression under the same meaning) for each training example.
- **Mixed Gaussian Recurrent Chain**: produce a series of diverse training data for each observed instance from its vicinity, i.e. a semantically-preserved continuous sub-space.

Next Step

- **Multilingual Scenario**: further study continuous semantic augmentation with the combination of large scale multilingual parallel and non-parallel corpora.

Summary

CsaNMT A novel data augmentation paradigm for NMT, which achieves state-of-the-art results

- **Tangential Contrastive Learning**: pre-train a semantic encoder, which forms an adjacency semantic region (i.e. the *vicinity*, cover adequate variants of literal expression under the same meaning) for each training example.
- **Mixed Gaussian Recurrent Chain**: produce a series of diverse training data for each observed instance from its vicinity, i.e. a semantically-preserved continuous sub-space.

Next Step

- **Multilingual Scenario**: further study continuous semantic augmentation with the combination of large scale multilingual parallel and non-parallel corpora.
- **Model Architecture**: develop continuous semantic augmentation as a pure data augmentation merged into the vanilla Transformer.

Q & A

Code: https://github.com/pemywei/csanmt

Contact: pemywei@gmail.com