Focused Concatenation for Context-Aware NMT

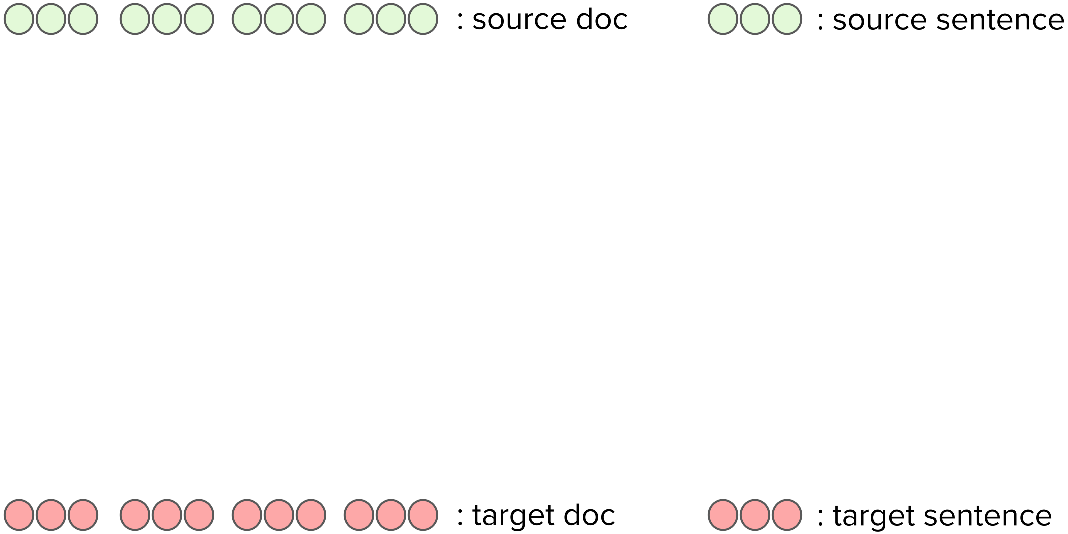
Lorenzo Lupo*, Marco Dinarelli*, Laurent Besacier*^

^ NAVER LABS Europe

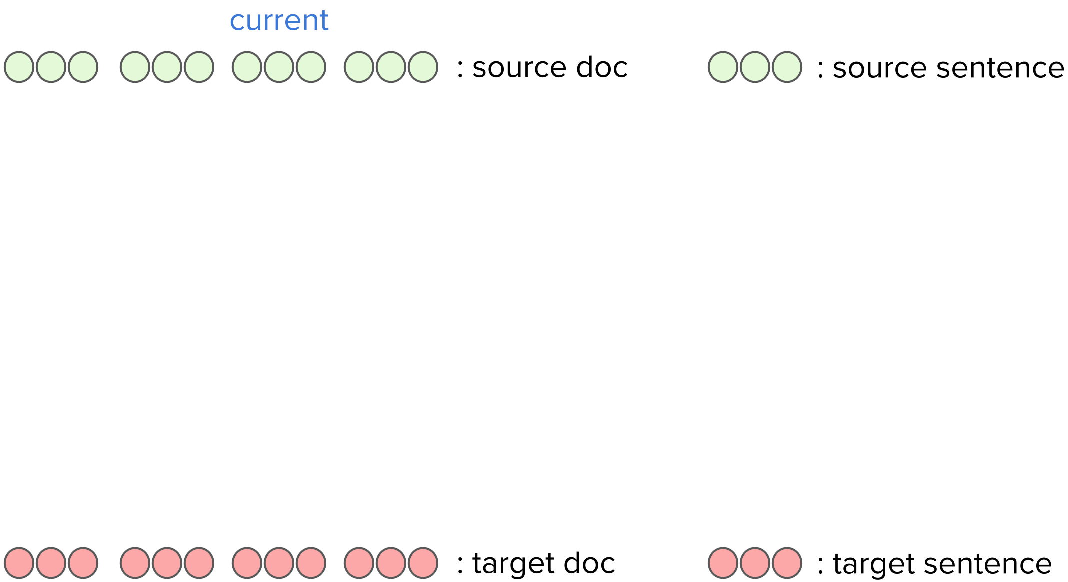
* Université Grenoble Alpes

*^ MIAI Grenoble Alpes Multidisciplinary Institute In Artificial Intelligence

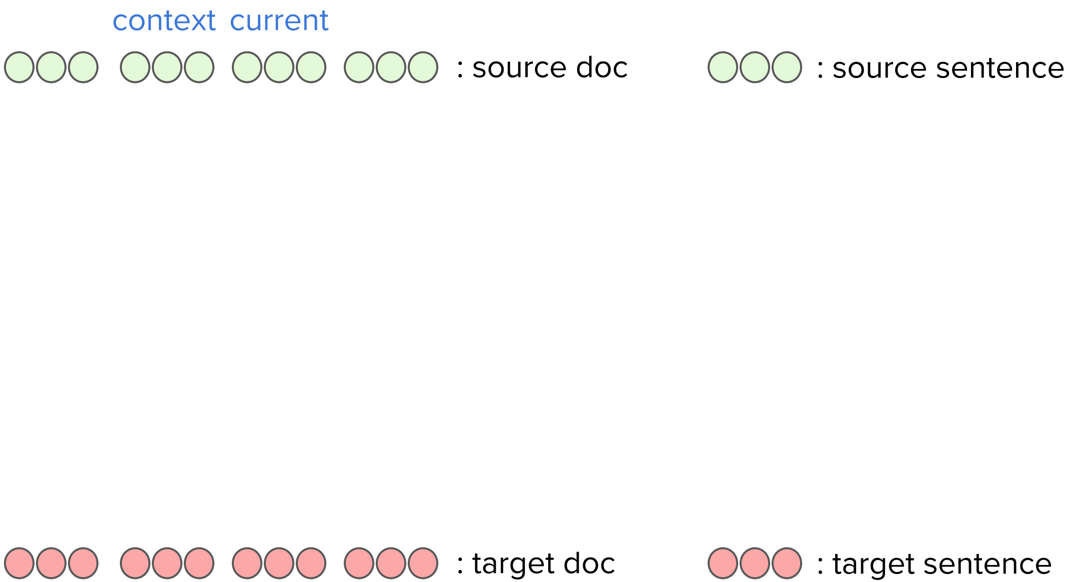
Context-aware NMT



Context-aware NMT

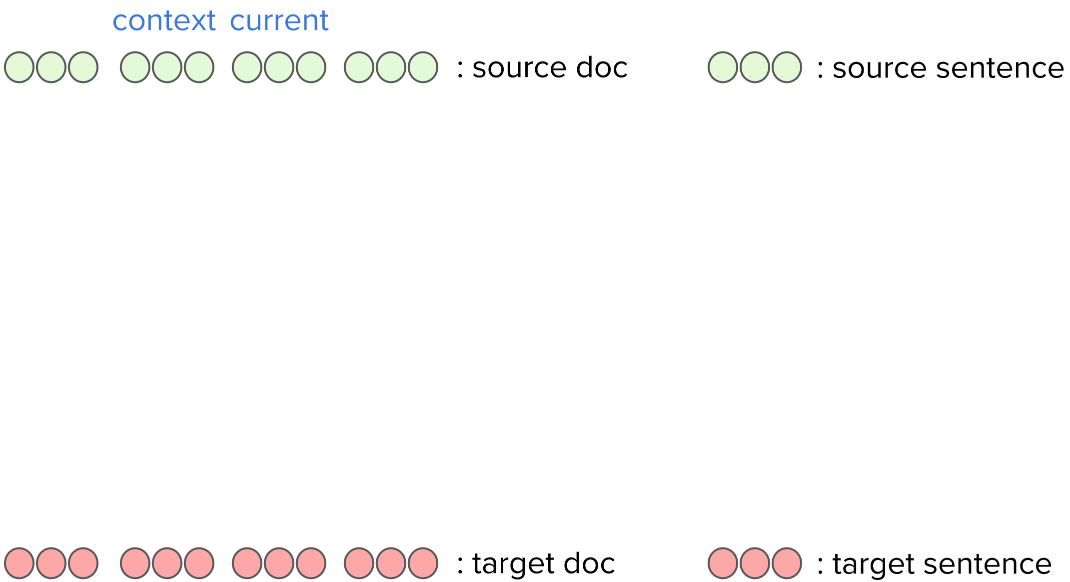


Context-aware NMT



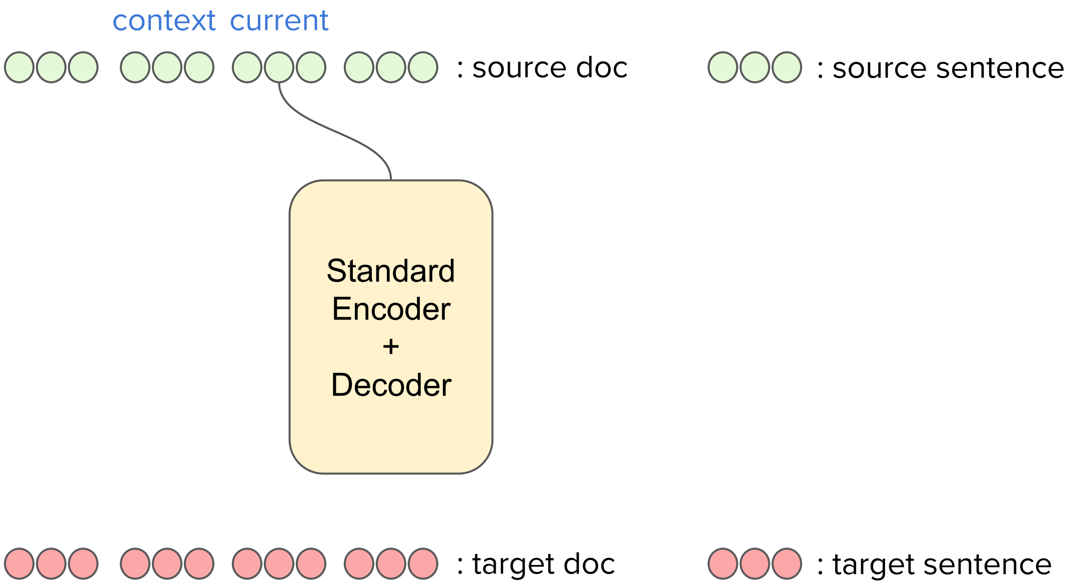
Context-aware NMT approaches

1. Multi-encoding approaches



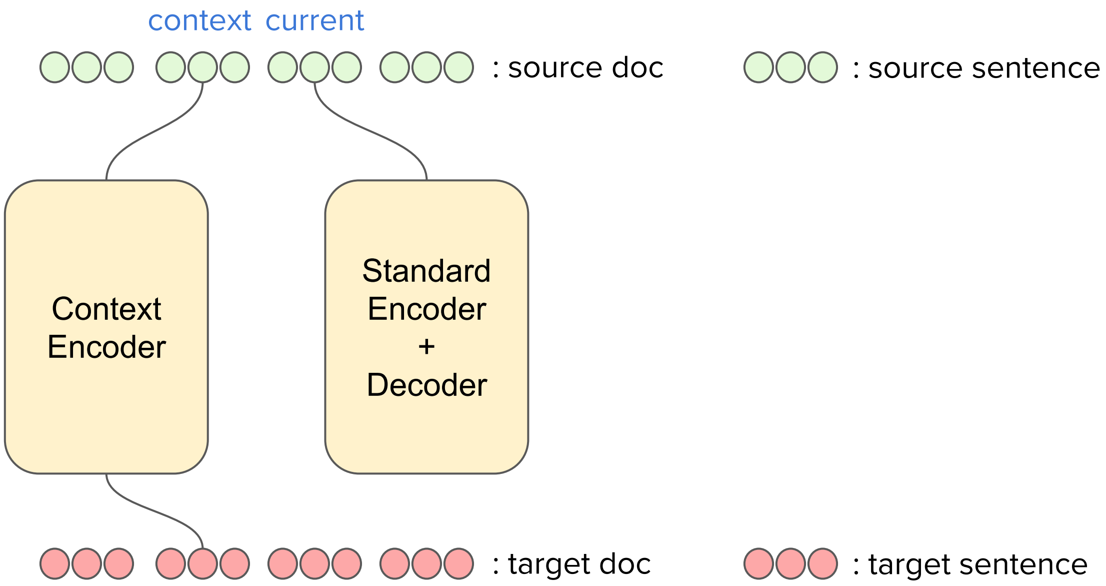
Context-aware NMT approaches

1. Multi-encoding approaches



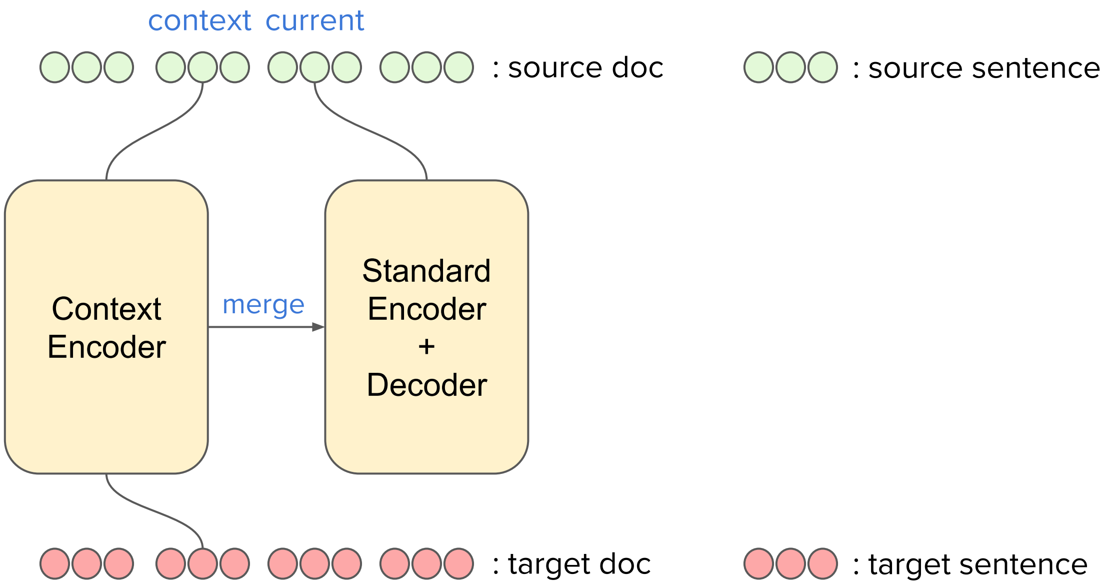
Context-aware NMT approaches

1. Multi-encoding approaches



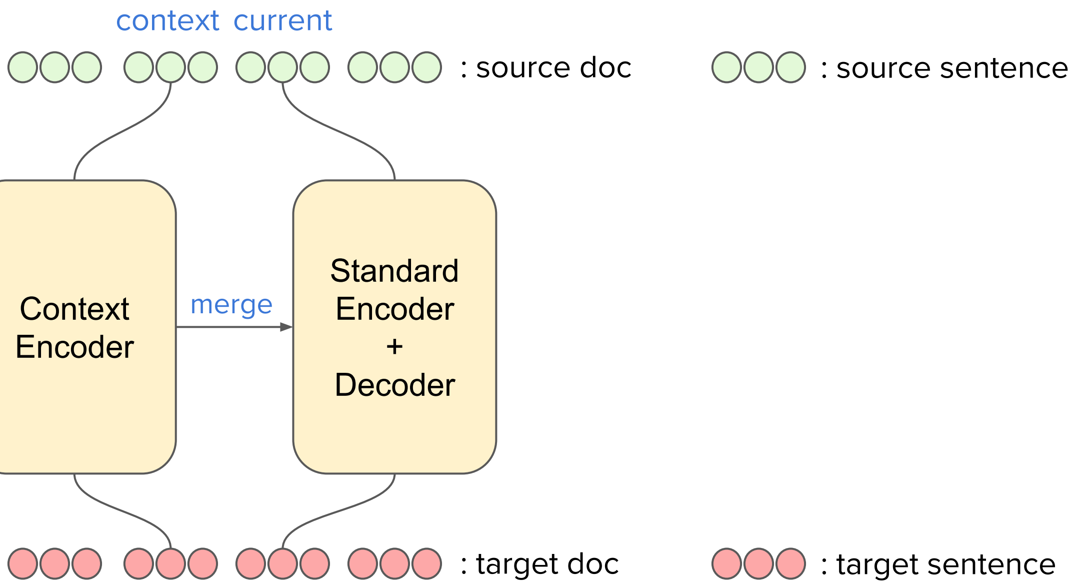
Context-aware NMT approaches

1. Multi-encoding approaches

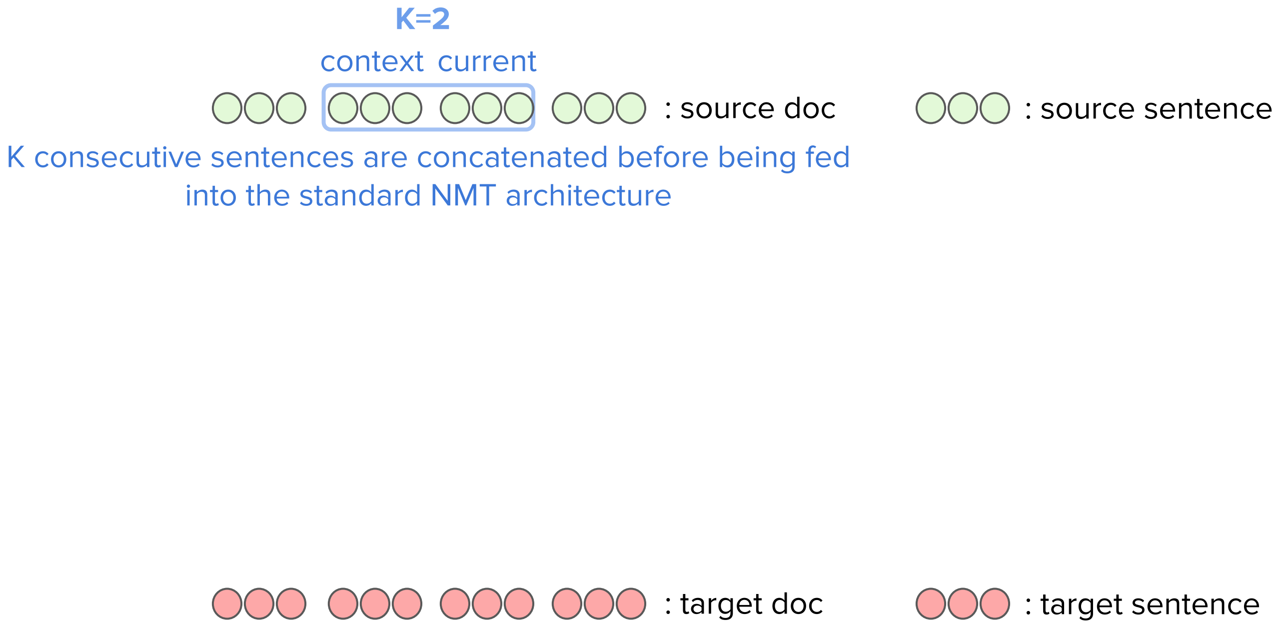


Context-aware NMT approaches

1. Multi-encoding approaches

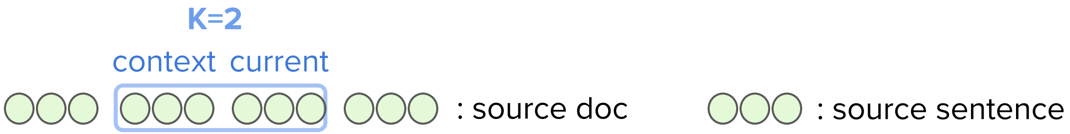


Context-aware NMT: the concatenation approach



Context-aware NMT: the concatenation approach

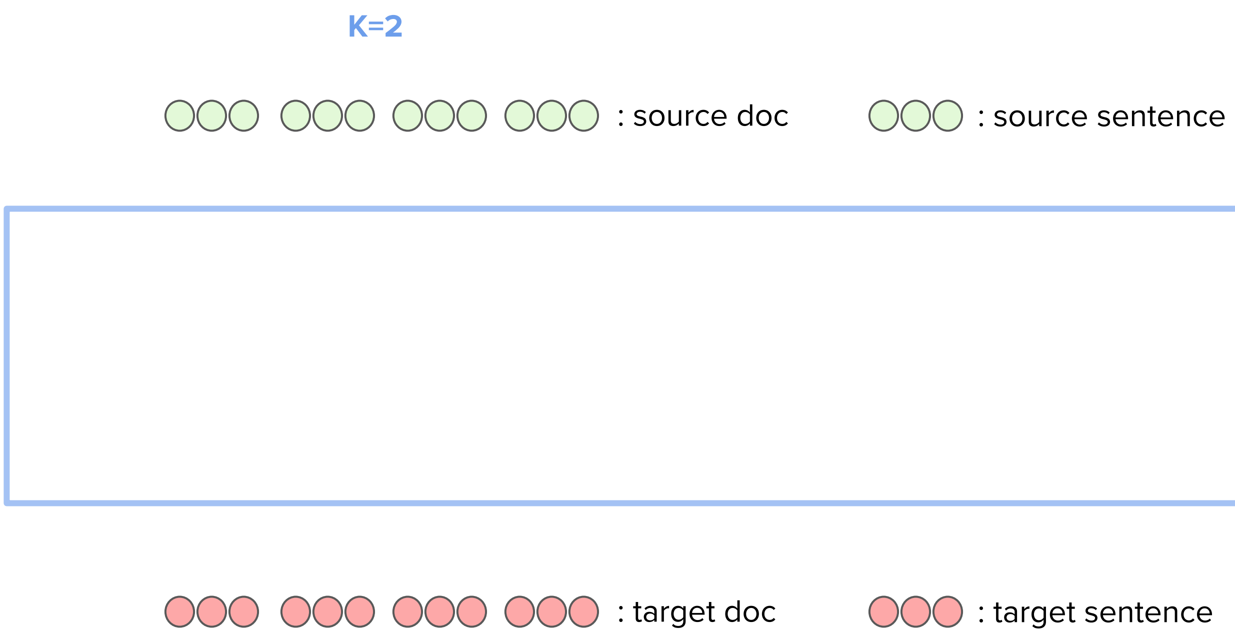
➔ SlidingKtoK



K consecutive sentences are concatenated before being fed into the standard NMT architecture

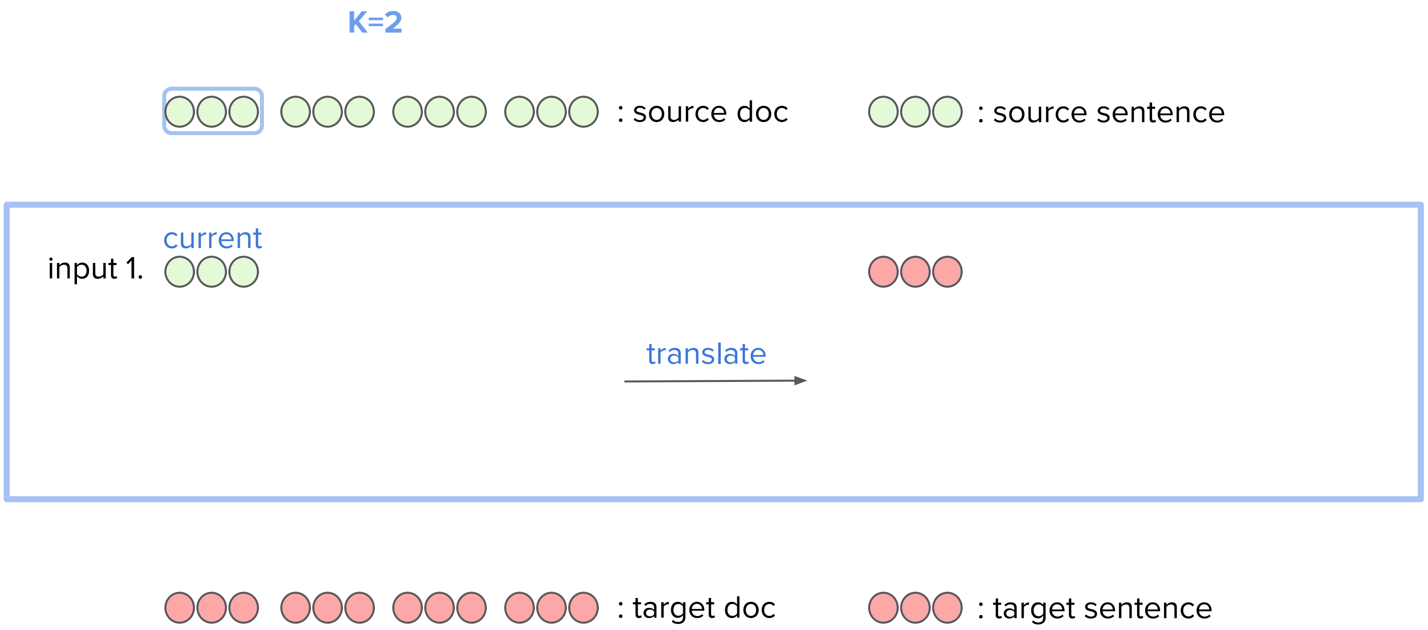
Context-aware NMT: the concatenation approach

➔ SlidingKtoK



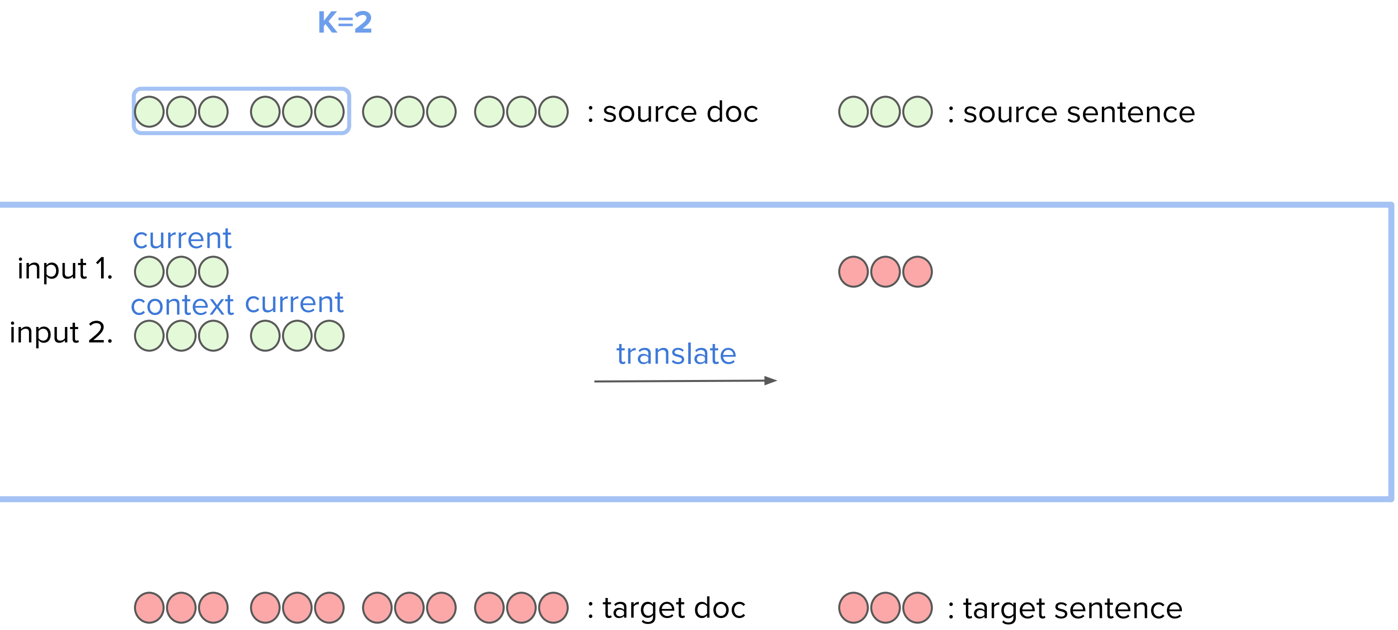
Context-aware NMT: the concatenation approach

➔ SlidingKtoK



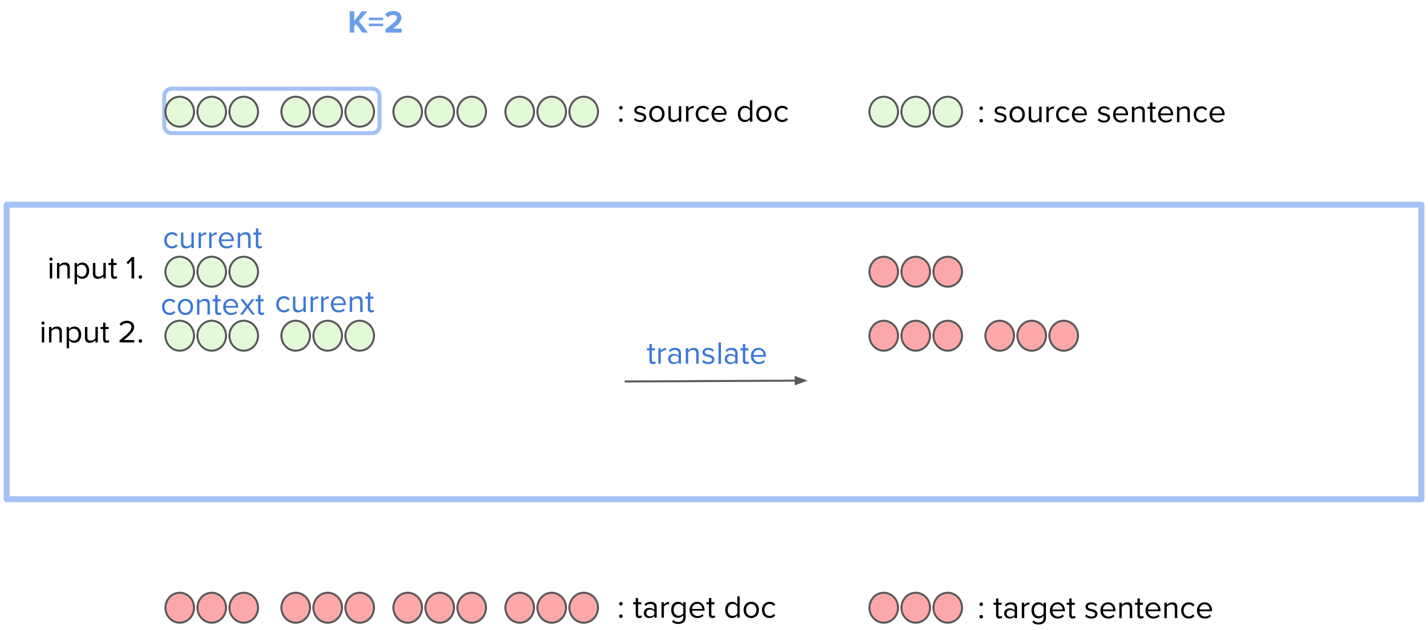
Context-aware NMT: the concatenation approach

➔ SlidingKtoK



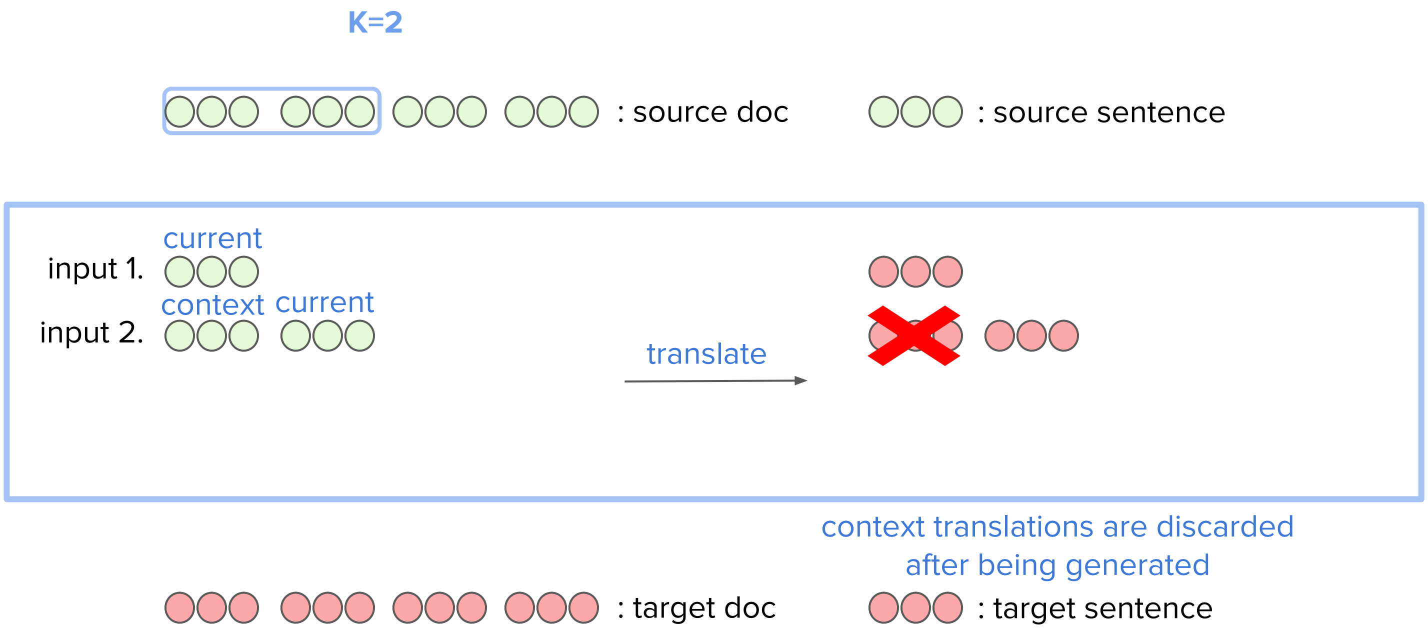
Context-aware NMT: the concatenation approach

➔ SlidingKtoK



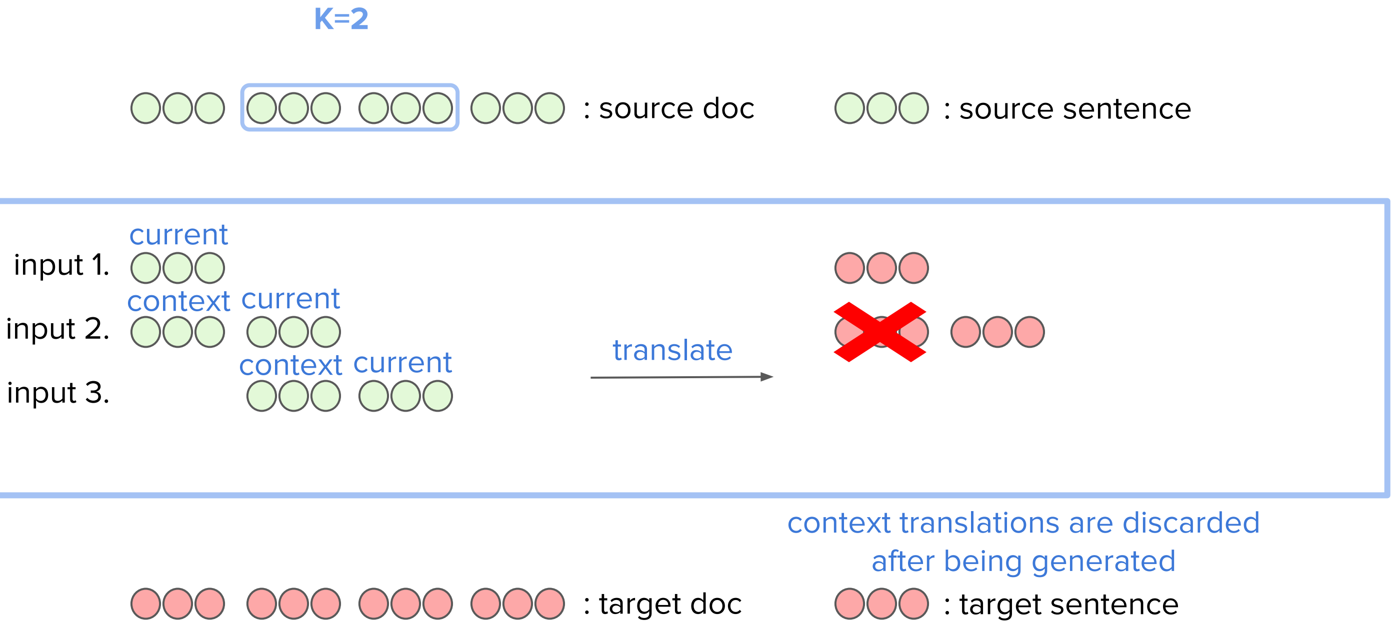
Context-aware NMT: the concatenation approach

➔ SlidingKtoK



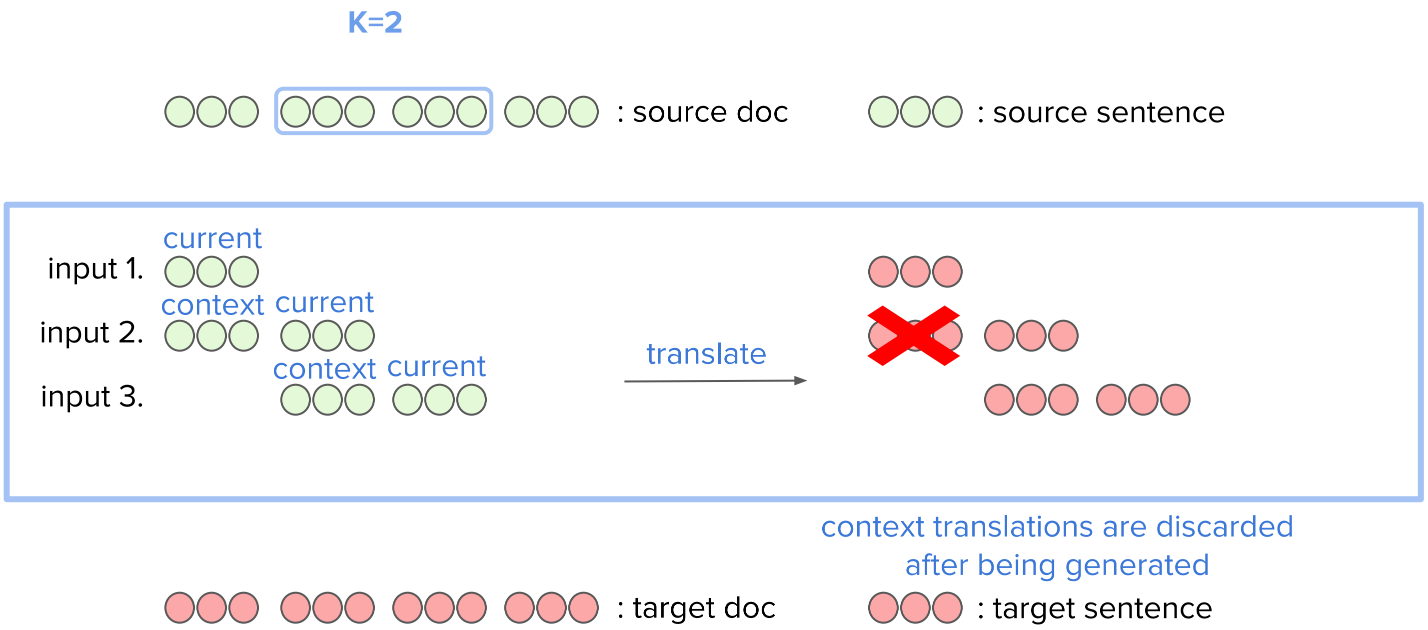
Context-aware NMT: the concatenation approach

➔ SlidingKtoK



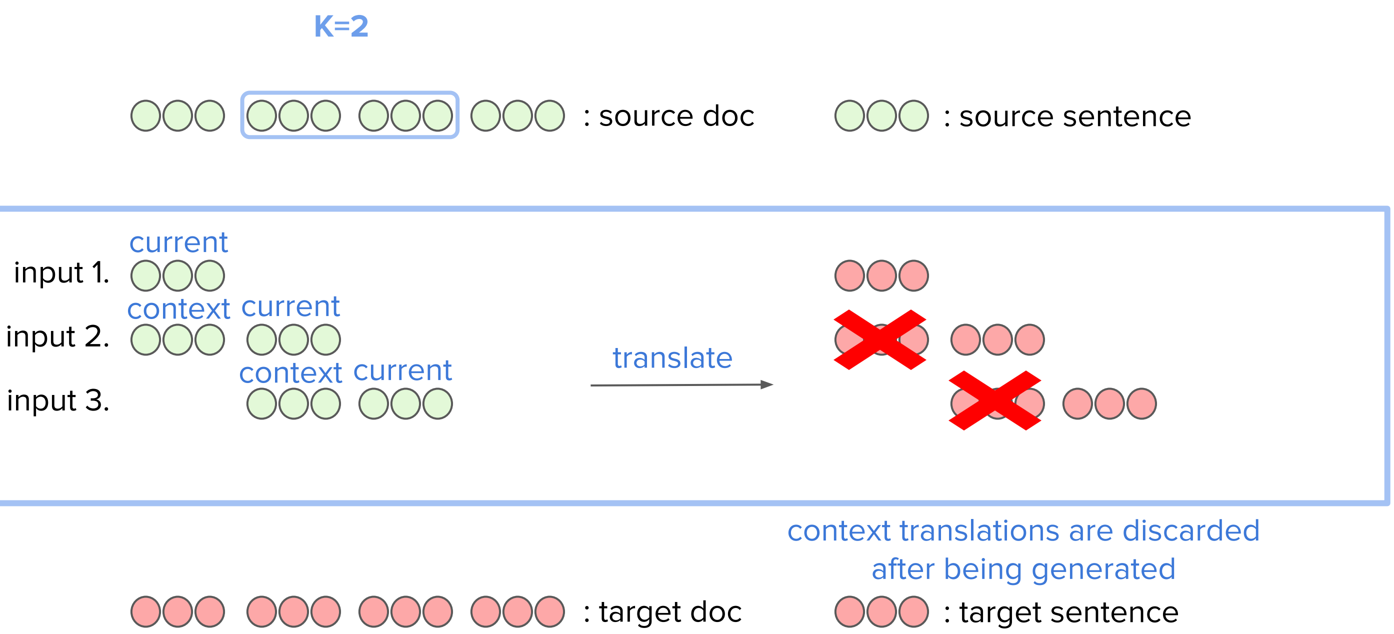
Context-aware NMT: the concatenation approach

➔ SlidingKtoK



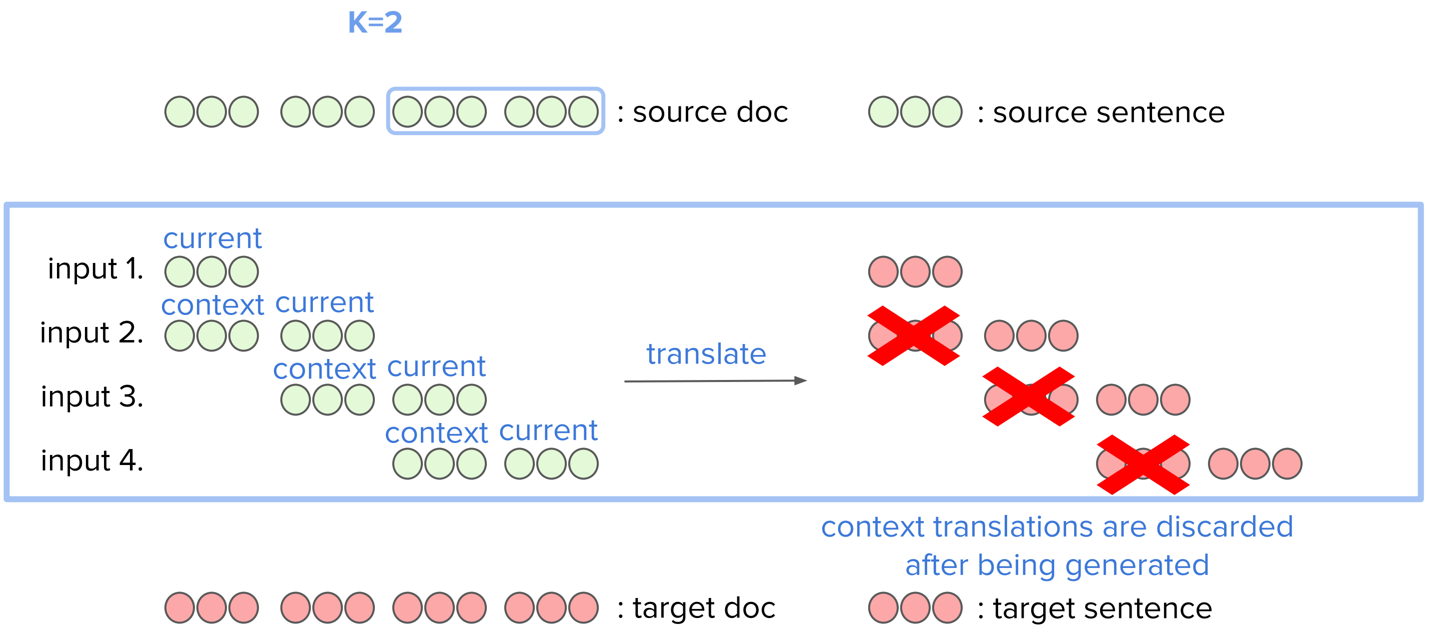
Context-aware NMT: the concatenation approach

➔ SlidingKtoK



Context-aware NMT: the concatenation approach

➔ SlidingKtoK



Context-aware NMT approaches

Strengths	Weaknesses
No extra learnable parameters added to the standard Transformer architecture.	

Context-aware NMT approaches

Strengths	Weaknesses
No extra learnable parameters added to the standard Transformer architecture.	
Since current and context sentences belong to the same sequence, **inter-sentential token contextualization** can be treated **in the same way as intra-sentential contextualization**.	

Context-aware NMT approaches

Strengths	Weaknesses
No extra learnable parameters added to the standard Transformer architecture.	**Attention can be *distracted* by context** instead of focusing on local syntactic structures [1], which are the most abundant, while context is important only for sparse tokens [2].
Since current and context sentences belong to the same sequence, **inter-sentential token contextualization** can be treated **in the same way as intra-sentential contextualization**.	

[1] Bao et al., 2021: *G-transformer for document-level machine translation.*
[2] Lupo et al., 2022: *Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models.*

Context-aware NMT approaches

Strengths	Weaknesses
No extra learnable parameters added to the standard Transformer architecture.	**Attention can be *distracted* by context** instead of focusing on local syntactic structures [1], which are the most abundant, while context is important only for sparse tokens [2].
Since current and context sentences belong to the same sequence, **inter-sentential token contextualization** can be treated **in the same way as intra-sentential contextualization**.	Even though we only keep the translation of the current sentence after generation, the standard translation **objective function is not focused on predictions of the current sentence**.

[1] Bao et al., 2021: *G-transformer for document-level machine translation.*
[2] Lupo et al., 2022: *Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models.*

Proposed approaches

Context-discounted objective

Training example

$$\boldsymbol{x}_K^j = \boldsymbol{x}^{j-K+1}\boldsymbol{x}^{j-K+2}...\boldsymbol{x}^{j-1}\boldsymbol{x}^j$$

$$\boldsymbol{y}_K^j = \boldsymbol{y}^{j-K+1}\boldsymbol{y}^{j-K+2}...\boldsymbol{y}^{j-1}\boldsymbol{y}^j$$

Conventional objective

$$\mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) = \sum_{t=1}^{|\boldsymbol{y}_K^j|} \log P(y_{K,t}^j | \boldsymbol{y}_{K,<t}^j, \boldsymbol{x}_K^j)$$

Context-discounted objective

Training example

$$\boldsymbol{x}_K^j = \boldsymbol{x}^{j-K+1}\boldsymbol{x}^{j-K+2}...\boldsymbol{x}^{j-1}\boldsymbol{x}^j$$

$$\boldsymbol{y}_K^j = \boldsymbol{y}^{j-K+1}\boldsymbol{y}^{j-K+2}...\boldsymbol{y}^{j-1}\boldsymbol{y}^j$$

Conventional objective

$$\mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) = \sum_{t=1}^{|\boldsymbol{y}_K^j|} \log P(y_{K,t}^j | \boldsymbol{y}_{K,<t}^j, \boldsymbol{x}_K^j)$$

Context-discounted objective

$$\begin{aligned}\mathcal{L}_{\text{CD}}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) &= \text{CD}\cdot\mathcal{L}_{context} + \mathcal{L}_{current} \\ &= \text{CD}\cdot\mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}_{K-1}^{j-1}) + \mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}^j)\end{aligned}$$

$$0 \leq \text{CD} < 1$$

Context-discounted objective

Training example

$$\boldsymbol{x}_K^j = \boldsymbol{x}^{j-K+1}\boldsymbol{x}^{j-K+2}...\boldsymbol{x}^{j-1}\boldsymbol{x}^j$$

$$\boldsymbol{y}_K^j = \boldsymbol{y}^{j-K+1}\boldsymbol{y}^{j-K+2}...\boldsymbol{y}^{j-1}\boldsymbol{y}^j$$

Conventional objective

$$\mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) = \sum_{t=1}^{|\boldsymbol{y}_K^j|} \log P(y_{K,t}^j | \boldsymbol{y}_{K,<t}^j, \boldsymbol{x}_K^j)$$

Context-discounted objective

$$\begin{aligned}\mathcal{L}_{\text{CD}}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) &= \text{CD}\cdot\mathcal{L}_{context} + \mathcal{L}_{current} \\ &= \text{CD}\cdot\mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}_{K-1}^{j-1}) + \mathcal{L}(\boldsymbol{x}_K^j, \boldsymbol{y}^j)\end{aligned}$$

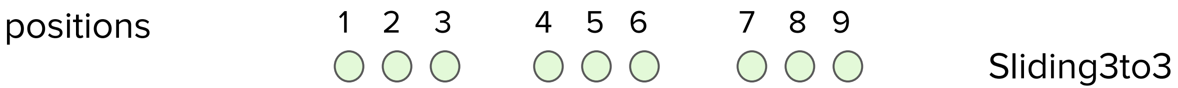
$$0 \leq \text{CD} < 1$$

➔ improve model focus on the current sentence;

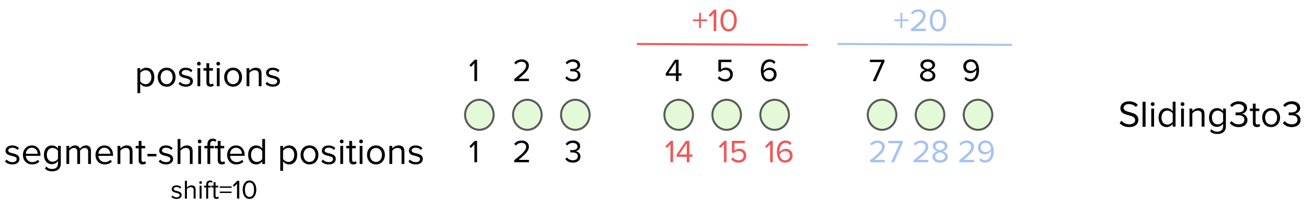
Segment-shifted position embeddings

Sliding3to3

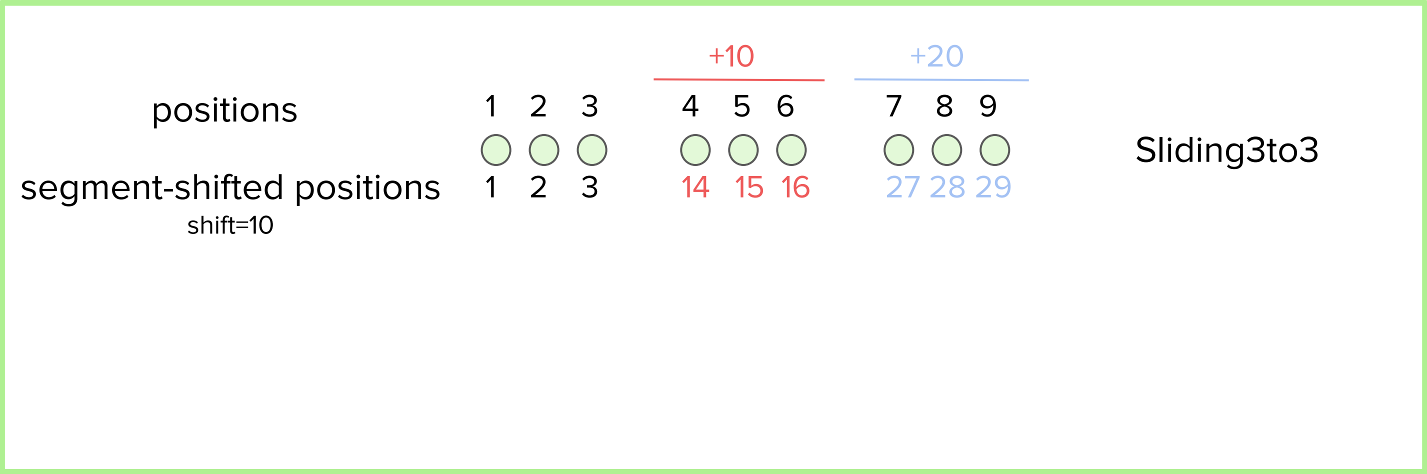
Segment-shifted position embeddings



Segment-shifted position embeddings

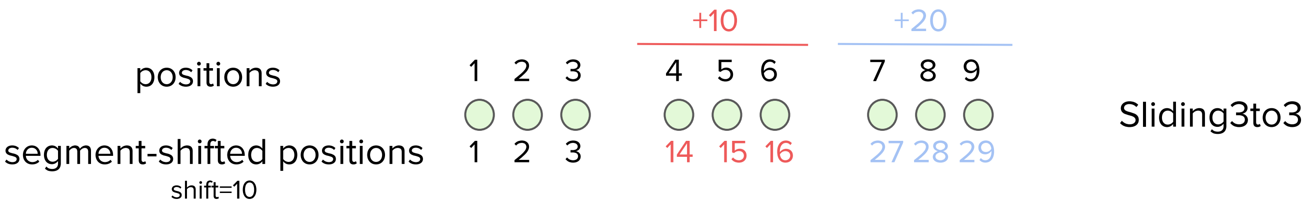


Segment-shifted position embeddings



- ➔ strengthen sentence boundaries;
- ➔ better distinguish between inter-sentential and intra-sentential discourse phenomena;
- ➔ no extra parameters (VS segment-embeddings like BERT).

Segment-shifted position embeddings



How big should be the shift?

- Average sentence length (in the corpus)
- Average sentence length (in the concatenated sequence)
- Big shift: shift >> average sentence length

➔ strengthen sentence boundaries;
➔ better distinguish between inter-sentential and intra-sentential discourse phenomena;
➔ no extra parameters (VS segment-embeddings like BERT).

Experiments

Experimental setup

Models
base: context-agnostic transformer-base
s4to4: sliding4to4 concatenation approach

Experimental setup

Models

base: context-agnostic transformer-base
s4to4: sliding4to4 concatenation approach

Data

English → Russian

- 6M sentence pairs from OpenSubtitles
- short documents of 4 sentences each

English → German

- 0.2M sentence pairs from IWSLT
- long documents of hundreds of sentences each

Experimental setup

Models

base: context-agnostic transformer-base
s4to4: sliding4to4 concatenation approach

Data

English → Russian

- 6M sentence pairs from OpenSubtitles
- short documents of 4 sentences each

English → German

- 0.2M sentence pairs from IWSLT
- long documents of hundreds of sentences each

Evaluation

BLEU on test set
+ **COMET**

Experimental setup

Models

base: context-agnostic transformer-base
s4to4: sliding4to4 concatenation approach

Data

English → Russian

- 6M sentence pairs from OpenSubtitles
- short documents of 4 sentences each

English → German

- 0.2M sentence pairs from IWSLT
- long documents of hundreds of sentences each

Evaluation

BLEU on test set
\+ **COMET**

Average translation quality metrics are scarcely sensitive to context-aware translation improvements, which affect a few words only. ➡ Targeted evaluation is necessary to appreciate model differences

Experimental setup

Models

base: context-agnostic transformer-base

s4to4: sliding4to4 concatenation approach

Data

English → Russian

- 6M sentence pairs from OpenSubtitles
- short documents of 4 sentences each

English → German

- 0.2M sentence pairs from IWSLT
- long documents of hundreds of sentences each

Evaluation

BLEU on test set

\+ **COMET**

\+ **Accuracy on targeted test sets** for the translation of discourse phenomena (**Disc.**):

- coreferential pronoun for En-De (ContraPro)
- deixis, lexical cohesion, noun phrase ellipsis, verb-phrase ellipsis for En-Ru (Voita)

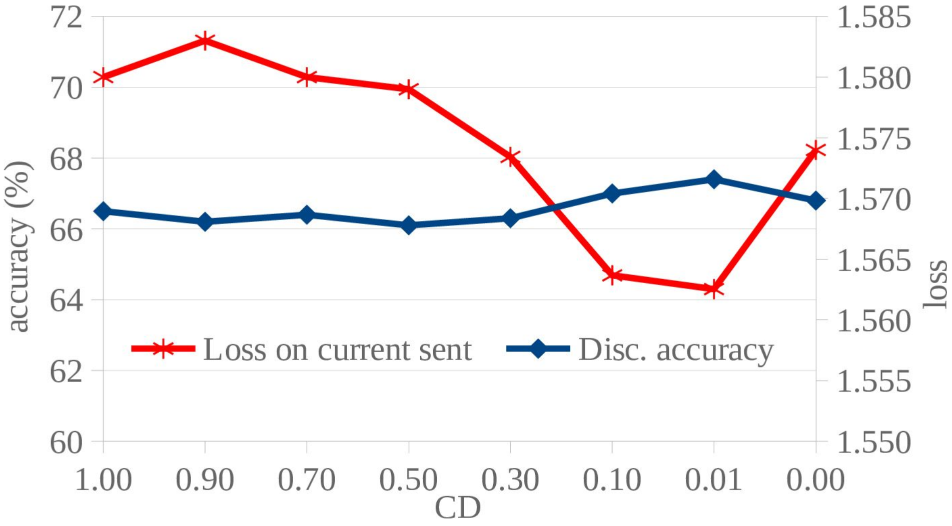
Average translation quality metrics are scarcely sensitive to context-aware translation improvements, which affect a few words only. ➡ Targeted evaluation is necessary to appreciate model differences

Preliminary analysis

$$\mathcal{L}_{\mathrm{CD}}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) = \mathrm{CD} \cdot \mathcal{L}_{context} + \mathcal{L}_{current}$$

Preliminary analysis

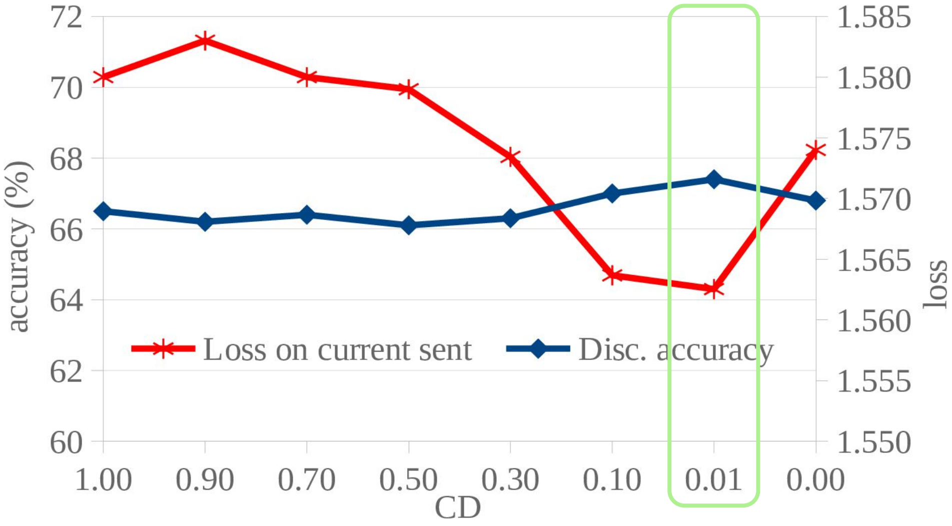
$$\mathcal{L}_{\text{CD}}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) = \text{CD}\cdot\mathcal{L}_{context} + \mathcal{L}_{current}$$



Evaluation of **En→Ru s4to4** trained with various levels of context discounting, ranging from 1 to 0.

Preliminary analysis

$$\mathcal{L}_{\text{CD}}(\boldsymbol{x}_K^j, \boldsymbol{y}_K^j) = \text{CD} \cdot \mathcal{L}_{context} + \mathcal{L}_{current}$$



Evaluation of **En→Ru s4to4** trained with various levels of context discounting, ranging from 1 to 0.

Main results

	En→Ru						
System	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	BLEU	COMET
base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
s4to4 + `CD`	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
s4to4 + shift + `CD`	85.76	**48.33***	**81.40**	**80.4***	**73.56***	32.37	0.334*

Main results

baselines:

	En→Ru						
System	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	BLEU	COMET
base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
s4to4 + CD	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
s4to4 + shift + CD	85.76	**48.33***	**81.40**	**80.4***	**73.56***	32.37	0.334*

Main results

	System	En→Ru Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	BLEU	COMET
baselines:	base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
	s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
our approaches	s4to4 + CD	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
	s4to4 + shift + CD	85.76	**48.33***	**81.40**	**80.4***	**73.56***	32.37	0.334*

Main results

	System	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	BLEU	COMET
				En→Ru				
baselines:	base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
	s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
our approaches	s4to4 + CD	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
	s4to4 + shift + CD	85.76	**48.33***	**81.40**	**80.4***	**73.56***	32.37	0.334*

Main results

accuracy (%) on discourse phenomena

	System	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	BLEU	COMET
				En→Ru				
baselines:	base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
	s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
our approaches	s4to4 + CD	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
	s4to4 + shift + CD	85.76	**48.33***	**81.40**	**80.4***	**73.56***	32.37	0.334*

The **accuracy on Disc.** is detailed on its left with the accuracy on each of the 4 subsets composing the targeted test set.

Main results

accuracy (%) on discourse phenomena

	System	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	BLEU	COMET
				En→Ru				
baselines:	base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
	s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
our approaches	s4to4 + CD	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
	s4to4 + shift + CD	85.76	**48.33***	**81.40**	**80.4***	**73.56***	32.37	0.334*

The **accuracy on Disc.** is detailed on its left with the accuracy on each of the 4 subsets composing the targeted test set.

Main results

accuracy (%) on discourse phenomena

			En→Ru				
System	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.	BLEU	COMET
base	50.00	45.87	51.80	27.00	46.64	31.98	0.321
s4to4	85.80	46.13	79.60	73.20	72.02	32.45	0.329
s4to4 + CD	**87.16***	46.40	81.00	78.20*	73.42*	32.37	0.328
s4to4 + shift + CD	85.76	**48.33***	**81.40**	**80.4***	**73.56***	32.37	0.334*
			En→De				
	d=1	d=2	d=3	d>3	Disc.	BLEU	COMET
base	32.89	43.97	47.99	70.58	37.27	29.63	0.546
s4to4	68.89	74.96	79.58	**87.78**	71.35	29.48	0.536
s4to4 + CD	**72.86***	75.96	80.10	84.38	74.31*	29.32	0.522
s4to4 + shift + CD	72.56*	**77.15***	**80.27**	86.65	**74.39***	29.20	0.528

baselines: base, s4to4

our approaches: s4to4 + CD, s4to4 + shift + CD

The **accuracy on Disc.** is detailed on its left with the accuracy on each of the 4 subsets composing the targeted test set.

Benchmarking

En→Ru

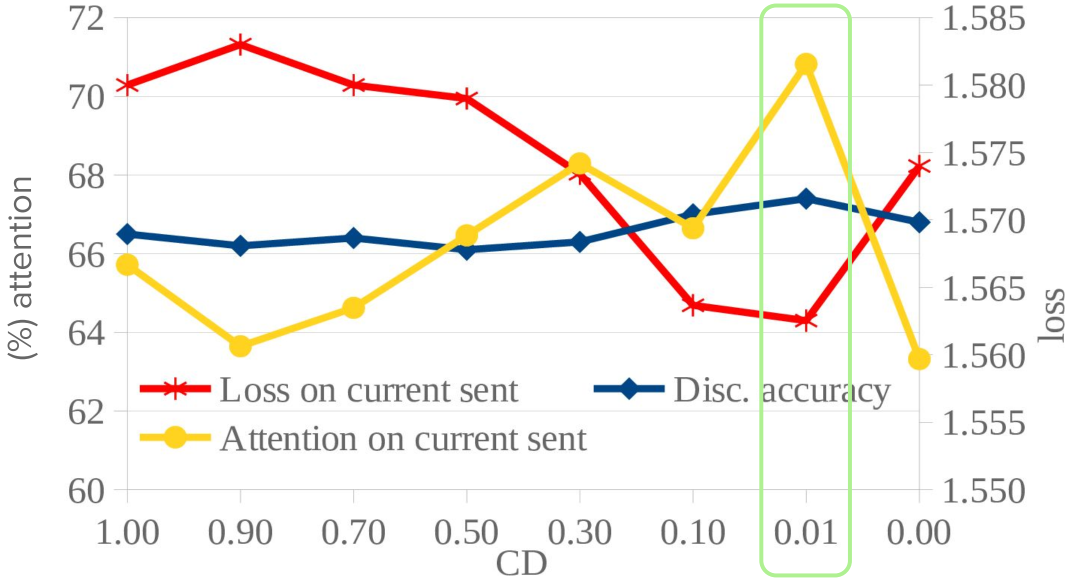
System	Deixis	Lex co.	Ell. inf	Ell. vp	Disc.
Chen et al. (2021)	62.30	47.90	64.90	36.00	55.61
Sun et al. (2022)	64.70	46.30	65.90	53.00	58.13
Zheng et al. (2020)	61.30	58.10	72.20	80.00	63.30
Kang et al. (2020)	79.20	62.00	71.80	80.80	73.46
Zhang et al. (2020)	**91.00**	46.90	78.20	**82.20**	**75.61**
s4to4 + shift + CD	85.76	**48.33**	**81.40**	80.40	73.56

En→De

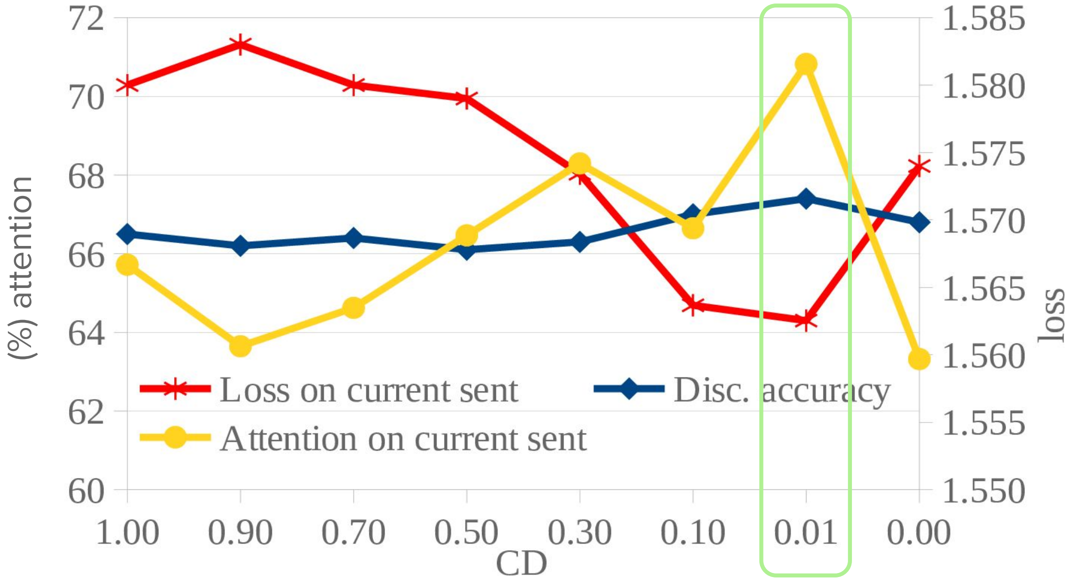
System	d=1	d=2	d=3	d>3	Disc.
Maruf et al. (2019)	34.70	46.40	51.10	70.10	39.15
Voita et al. (2018)	39.00	48.00	54.00	66.00	42.55
Stojanovski and Fraser (2019)	53.00	46.00	50.00	71.00	52.55
Lupo et al. (2022)	56.50	44.90	48.70	73.30	54.98
Müller et al. (2018)	58.00	55.00	55.00	75.00	58.13
s4to4 + shift + CD	**72.56**	**77.15**	**80.27**	**86.65**	**74.39**

Analysis

Impact on the distribution of attention weights

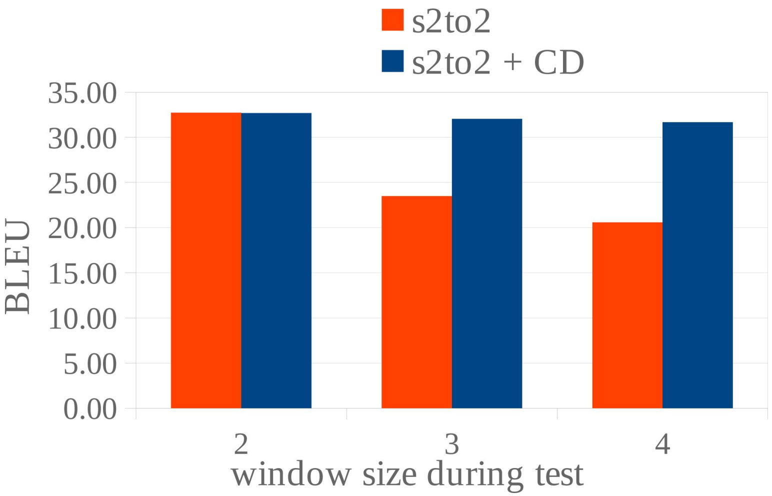


Self-attention gets more focused



System	Attn entropy
s4to4	2.293
s4to4 + CD	2.276
s4to4 + shift + CD	**2.251**

Concatenation becomes robust to context size



Takeaways

A sliding windows approach trained with a **context-discounted objective** function

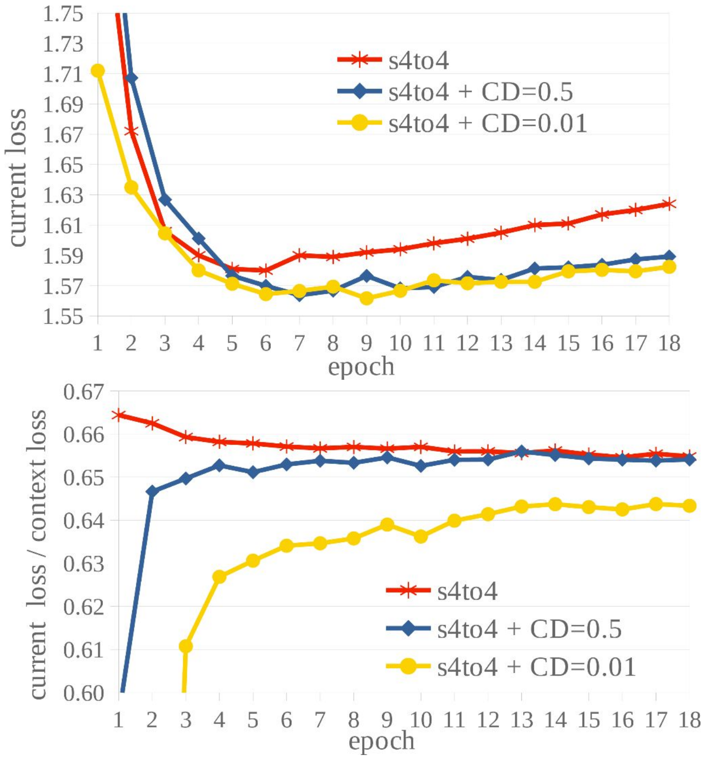
1. Performs **better on** the disambiguation of inter-sentential **discourse phenomena**;
2. **Improves predictions** of the current reference;
3. Learn self-attention modules that are **less distracted** by context;
4. Is **more robust to context** sizes unseen during training.

Segment-shifted position embeddings further help focusing attention and slightly improve performance.

Thank you for listening!

Link to *Focused Concatenation for Context-Aware Neural Machine Translation*

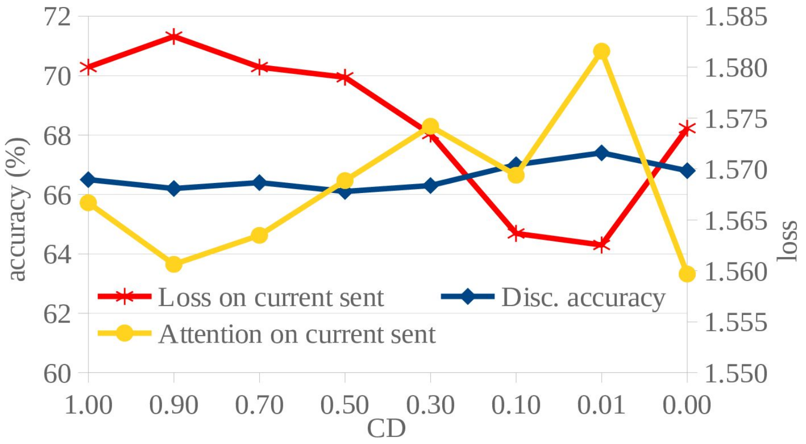
Analysis of context discounting



Our empirical analysis on concatenation models trained with the context-discounted objective shows that **context discounting enables**:

1. **better predictions of the current target sentence** (lower validation loss), both absolutely (top plot) and relatively to the quality of the prediction of target context (bottom plot);

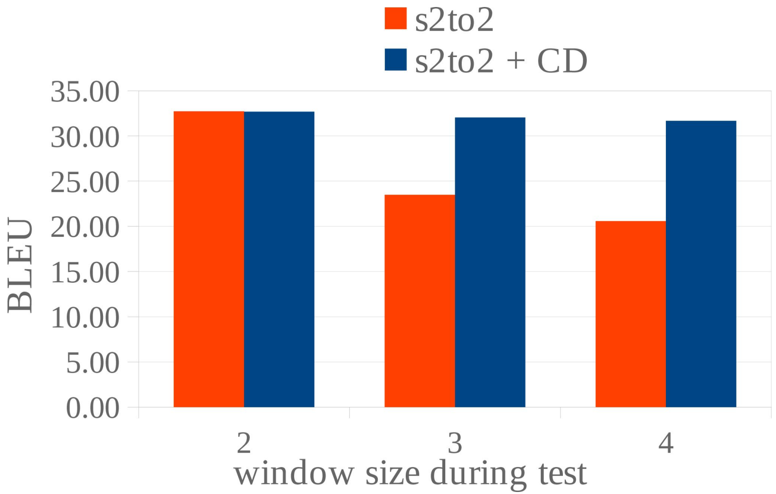
Analysis of context discounting



Our empirical analysis on concatenation models trained with the context-discounted objective shows that **context discounting enables**:

1. **better predictions of the current target sentence** (lower validation loss), both absolutely (top plot) and relatively to the quality of the prediction of target context (bottom plot);
2. **increased focus of self-attention on the current sentence**: the stronger the context discounting the stronger **the average portion of attention that is focused on the current sentence** from tokens belonging to the current sentence;

Analysis of context discounting



Our empirical analysis on concatenation models trained with the context-discounted objective shows that **context discounting enables**:

1. **better predictions of the current target sentence** (lower validation loss), both absolutely (top plot) and relatively to the quality of the prediction of target context (bottom plot);
2. **increased focus of self-attention on the current sentence**: the stronger the context discounting the stronger the average portion of attention that is focused on the current sentence from tokens belonging to the current sentence;
3. **robustness** of concatenation models to windows sizes unseen during training.

Analysis of segment-shifted position embeddings

We also analysed how the distribution of attention weights changes when adding segment-shifted position embeddings, finding that:

System	Attn entropy
s4to4	2.293
s4to4 + CD	2.276
s4to4 + shift + CD	**2.251**

Analysis of segment-shifted position embeddings

System	Attn entropy
s4to4	2.293
s4to4 + CD	2.276
s4to4 + shift + CD	**2.251**

We also analysed how the distribution of attention weights changes when adding segment-shifted position embeddings, finding that:

1. **Average entropy of self and cross-attention weights decreases** with the help of context discounting and segment-shifted positions.

Analysis of segment-shifted position embeddings

System	Shift	Disc.	BLEU
s4to4 + shift + CD	100	73.68	32.41
s4to4 + shift + CD	avg-sequence	73.38	32.37
s4to4 + shift + CD	avg-corpus	**73.97**	32.45

We also analysed how the distribution of attention weights changes when adding segment-shifted position embeddings, finding that:

1. **Average entropy of self and cross-attention weights decreases** with the help of context discounting and segment-shifted positions.

Finally, we performed two ablation studies:

2. a comparison between models adopting **different values of segment shifting**. No significant differences (p > 0.05);

Analysis of segment-shifted position embeddings

	En→Ru		En→De	
System	Disc.	BLEU	Disc.	BLEU
s4to4 + shift + CD	73.56	32.45	**74.39**	29.20
s4to4 + lrn + CD	**73.68**	32.45	72.14	28.35
s4to4 + sin + CD	73.48	32.53	73.88	29.23

We also analysed how the distribution of attention weights changes when adding segment-shifted position embeddings, finding that:

1. **Average entropy of self and cross-attention weights decreases** with the help of context discounting and segment-shifted positions.

Finally, we performed two ablation studies:

2. a comparison between models adopting **different values of segment shifting**. No significant differences ($p > 0.05$);
3. a **comparison with learned segment embeddings** and **sinusoidal segment embeddings**. No significant differences ($p > 0.05$), except for s4to4+lrn+CD on En→De.