Challenges and Remedies for Context-Aware Neural Machine Translation

Lorenzo Lupo

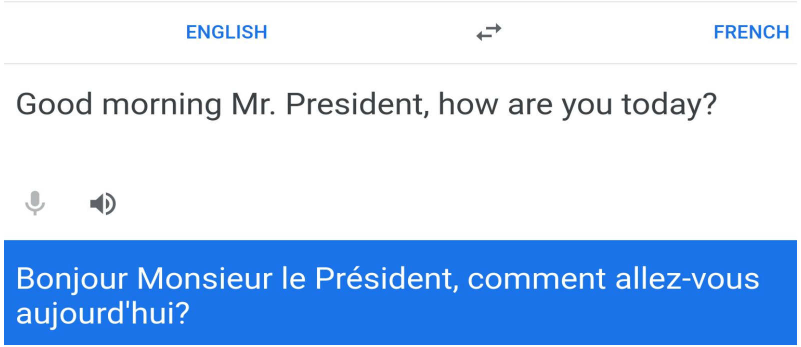
Outline

1. Introduction
2. Multi-encoding approaches
   a. Lupo, L., Dinarelli, M. and Besacier, L., **Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder NMT**, ACL 2022.
3. Concatenation approaches
   a. Lupo, L., Dinarelli, M. and Besacier, L., **Focused Concatenation for Context-Aware NMT**, WMT 2022.
   b. Lupo, L., Dinarelli, M. and Besacier, L., **Encoding Sentence Position in Context-Aware NMT with Concatenation**, Insights 2023.
4. Conclusions

Outline

1. Introduction
2. Multi-encoding approaches
    a. Lupo, L., Dinarelli, M. and Besacier, L., **Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder NMT**, ACL 2022.
3. Concatenation approaches
    a. Lupo, L., Dinarelli, M. and Besacier, L., **Focused Concatenation for Context-Aware NMT**, WMT 2022.
    b. Lupo, L., Dinarelli, M. and Besacier, L., **Encoding Sentence Position in Context-Aware NMT with Concatenation**, Insights 2023.
4. Conclusions

Context-aware NMT: why?



trial date: November 2022

Context-aware NMT: why?

ENGLISH ⇄ FRENCH

Good morning Mr. President, how are you today?

Bonjour Monsieur le Président, comment allez-vous aujourd'hui?

trial date: November 2022

Context-aware NMT: why?

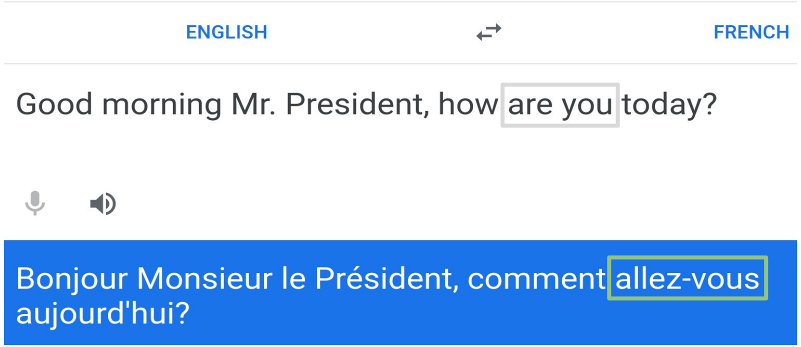
ENGLISH ⇄ FRENCH

Good morning Mr. President, how are you today?

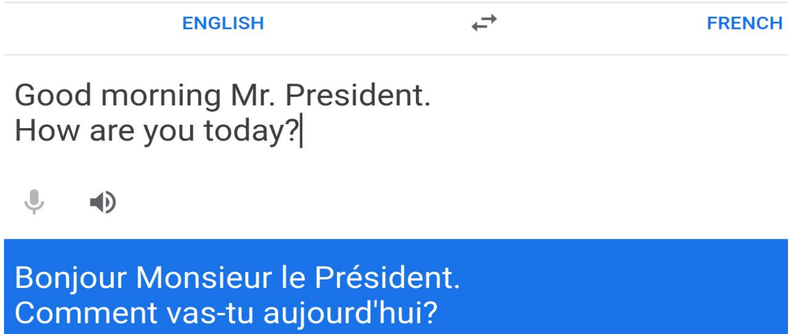
Bonjour Monsieur le Président, comment allez-vous aujourd'hui?

trial date: November 2022

Context-aware NMT: why?



trial date: November 2022



Context-aware NMT: why?

ENGLISH ⇄ FRENCH

Good morning Mr. President, how are you today?

Bonjour Monsieur le Président, comment allez-vous aujourd'hui?

trial date: November 2022

ENGLISH ⇄ FRENCH

Good morning Mr. President.
How are you today?

Bonjour Monsieur le Président.
Comment vas-tu aujourd'hui?

Context-aware NMT: why?

Research showed that a crucial challenge for neural machine translation (NMT) **to reach human quality** is the ability to **exploit inter-sentential context** - the preceding or following sentences in the same document [Läubli et al., 2018; Toral et al., 2018; Castilho et al., 2020]

Context-aware NMT: what?

Source document $X = \{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^{|X|}\}$

Target document $Y = \{\boldsymbol{y}^{\mathbf{1}}, \boldsymbol{y}^{\mathbf{2}}, ..., \boldsymbol{y}^{|Y|}\}$

Context-aware NMT: what?

Problem

$$P_\theta(Y|X) = \prod_{j=1}^{|X|} \prod_{t=1}^{|\boldsymbol{y}|} P_\theta(y_t^j | \boldsymbol{y}_{<t}^j, \boldsymbol{x}^j, context)$$

Source document

$$X = \{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^{|X|}\}$$

Target document

$$Y = \{\boldsymbol{y}^{\mathbf{1}}, \boldsymbol{y}^{\mathbf{2}}, ..., \boldsymbol{y}^{|Y|}\}$$

Context-aware NMT: what?

Problem

$$P_\theta(Y|X) = \prod_{j=1}^{|X|} \prod_{t=1}^{|\boldsymbol{y}|} P_\theta(y_t^j | \boldsymbol{y}_{<t}^j, \boldsymbol{x}^j, \boxed{context})$$

- **All the available sentences** in the parallel document.

Context-aware NMT: what?

Problem

$$P_\theta(Y|X) = \prod_{j=1}^{|X|} \prod_{t=1}^{|\boldsymbol{y}|} P_\theta(y_t^j | \boldsymbol{y}_{<t}^j, \boldsymbol{x}^j, \boxed{context})$$

- **All the available sentences** in the parallel document.
- **The parallel document and its meta-data**:
    - author’s information;
    - date of the writing;
    - domain of the writing;
    - visual context.

Context-aware NMT: what?

Problem

$$P_\theta(Y|X) = \prod_{j=1}^{|X|} \prod_{t=1}^{|\boldsymbol{y}|} P_\theta(y_t^j|\boldsymbol{y}_{<t}^j, \boldsymbol{x}^j, \boxed{context})$$

- **All the available sentences** in the parallel document.
- **The parallel document and its meta-data**:
    - author's information;
    - date of the writing;
    - domain of the writing;
    - visual context.
- **A few neighbouring sentences.**

Context-aware NMT: what?

Problem

$$P_\theta(Y|X) = \prod_{j=1}^{|X|} \prod_{t=1}^{|\boldsymbol{y}|} P_\theta(y_t^j | \boldsymbol{y}_{<t}^j, \boldsymbol{x}^j, \boxed{context})$$

- **All the available sentences** in the parallel document.
- **The parallel document and its meta-data**:
  - author's information;
  - date of the writing;
  - domain of the writing;
  - visual context.
- **A few neighbouring sentences.**

Most existing approaches use **a few preceding sentences** [Maruf et al., 2021], where **most of the disambiguating information** is present [Castilho et al., 2020] .

Context-aware NMT: what?

Problem

$$P_\theta(Y|X) = \prod_{j=1}^{|X|} \prod_{t=1}^{|\boldsymbol{y}|} P_\theta(y_t^j | \boldsymbol{y}_{<t}^j, \boldsymbol{x}^j, \boxed{context})$$

Training corpus

$$\mathcal{C} = \{(X^1, Y^1), (X^2, Y^2), ..., (X^D, Y^D)\}$$

Training objective

$$\underset{\theta}{\operatorname{argmin}} \sum_{d \in \mathcal{C}} -\log P_\theta(Y^d|X^d)$$

Context-aware NMT: what?

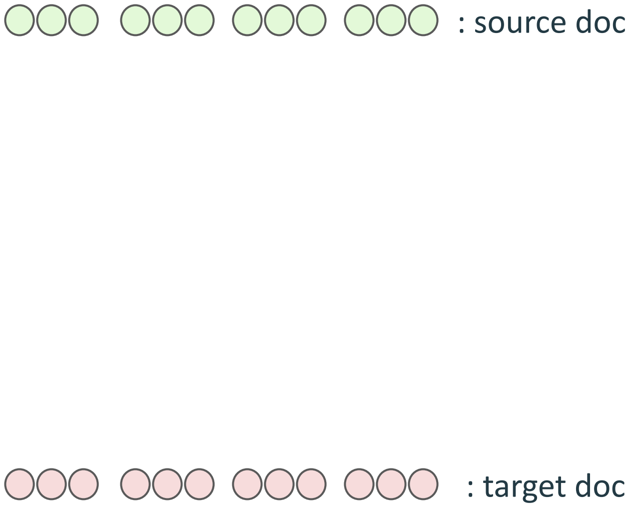
Source document	$X = \{\boldsymbol{x}^1, \boldsymbol{x}^2, ..., \boldsymbol{x}^{\|X\|}\}$
Target document	$Y = \{\boldsymbol{y}^{\mathbf{1}}, \boldsymbol{y}^{\mathbf{2}}, ..., \boldsymbol{y}^{\|Y\|}\}$
Problem	$P_\theta(Y\|X) = \prod_{j=1}^{\|X\|} \prod_{t=1}^{\|\boldsymbol{y}\|} P_\theta(y_t^j \| \boldsymbol{y}_{<t}^j, Y_{<j}, X)$
Training corpus	$\mathcal{C} = \{(X^1, Y^1), (X^2, Y^2), ..., (X^D, Y^D)\}$
Training objective	$\underset{\theta}{\operatorname{argmin}} \sum_{d \in \mathcal{C}} - \log P_\theta(Y^d \| X^d)$

Context-aware NMT: how? [Kim et al.,2019]

Concatenation	Multi-encoding

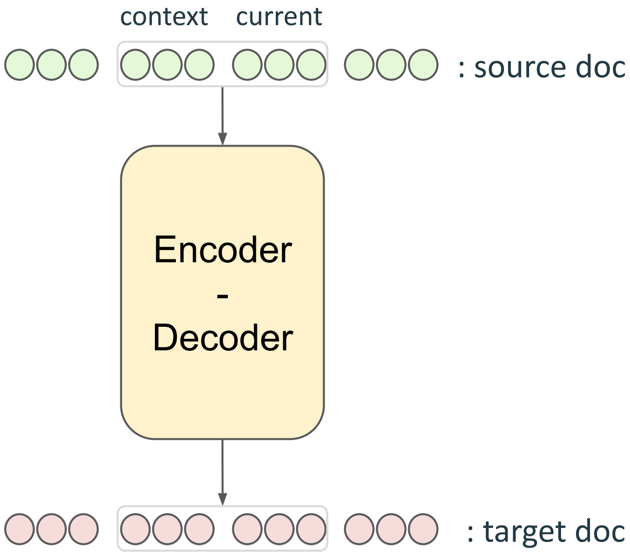
Context-aware NMT: how? [Kim et al.,2019]

Concatenation	Multi-encoding



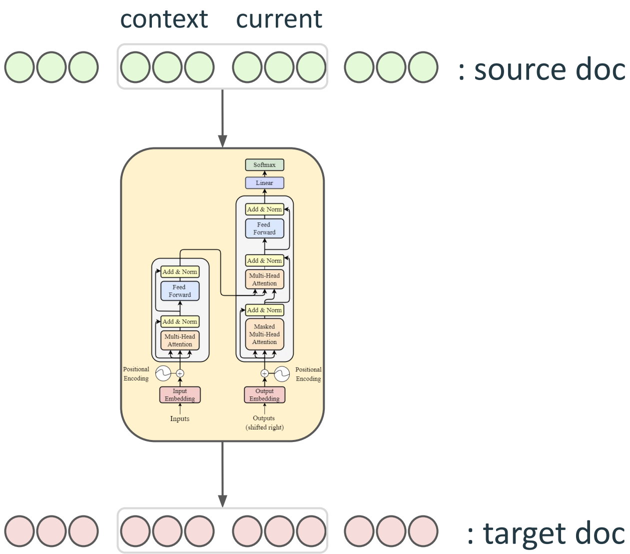
Context-aware NMT: how? [Kim et al.,2019]

Concatenation	Multi-encoding



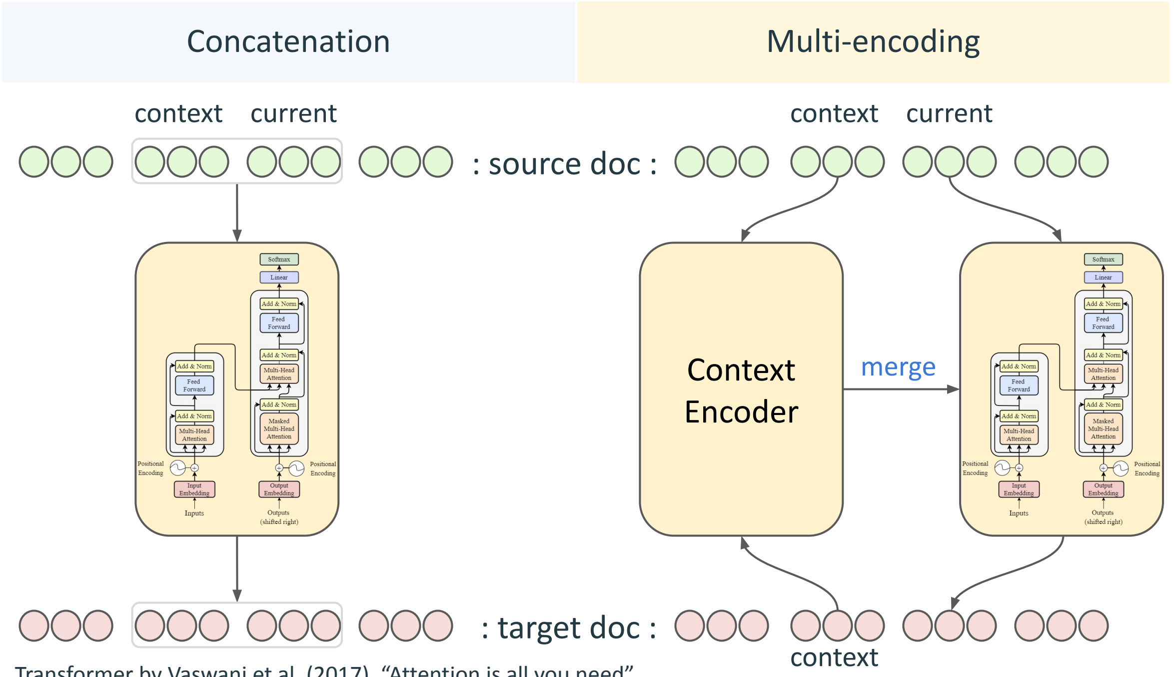
Context-aware NMT: how? [Kim et al.,2019]

Concatenation	Multi-encoding



Transformer by Vaswani et al. (2017), “Attention is all you need”.

Context-aware NMT: how? [Kim et al.,2019]



Transformer by Vaswani et al. (2017), “Attention is all you need”.

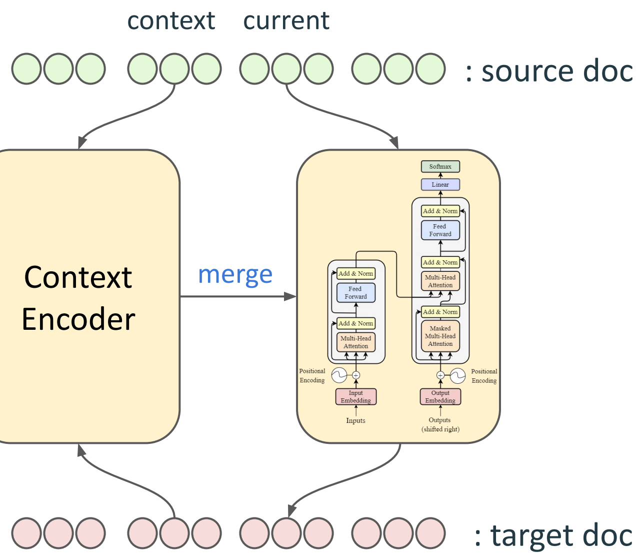
Objectives

1. **Identify challenges** in both multi-encoding and concatenation approaches.
2. **Propose remedies** to tackle the challenges identified.
3. **Improve understanding** through the analysis of the proposed solutions.

Outline

1. Introduction
2. Multi-encoding approaches
    a. Lupo, L., Dinarelli, M. and Besacier, L., **Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder NMT**, ACL 2022.
3. Concatenation approaches
    a. Lupo, L., Dinarelli, M. and Besacier, L., **Focused Concatenation for Context-Aware NMT**, WMT 2022.
    b. Lupo, L., Dinarelli, M. and Besacier, L., **Encoding Sentence Position in Context-Aware NMT with Concatenation**, Insights 2023.
4. Conclusions

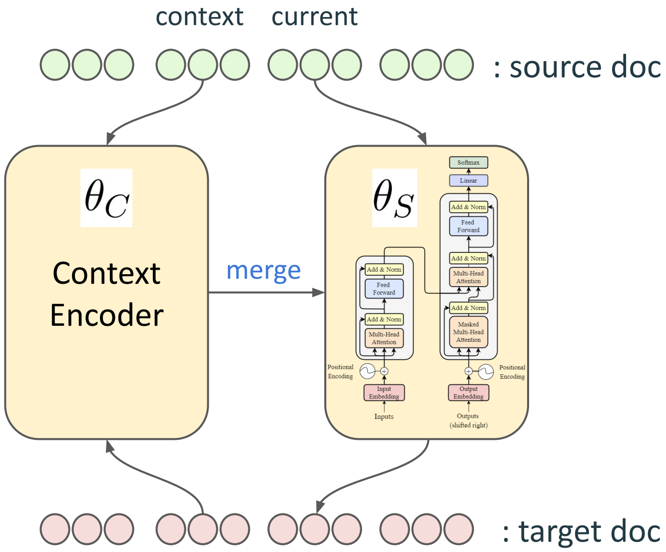
Multi-encoding approaches



Transformer by Vaswani et al. (2017), "Attention is all you need".

Multi-encoding approaches

Trainable parameters: $\Theta = [\theta_S; \theta_C]$

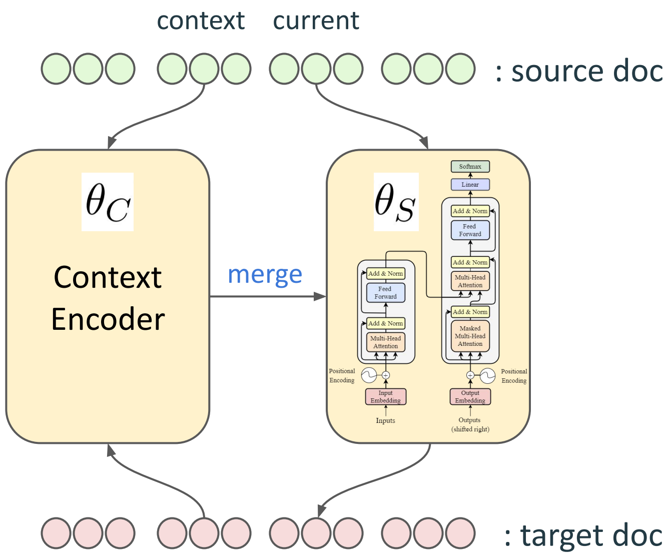


Transformer by Vaswani et al. (2017), “Attention is all you need”.

Multi-encoding approaches

contextual parameters

Trainable parameters: $\Theta = [\theta_S; \boxed{\theta_C}]$

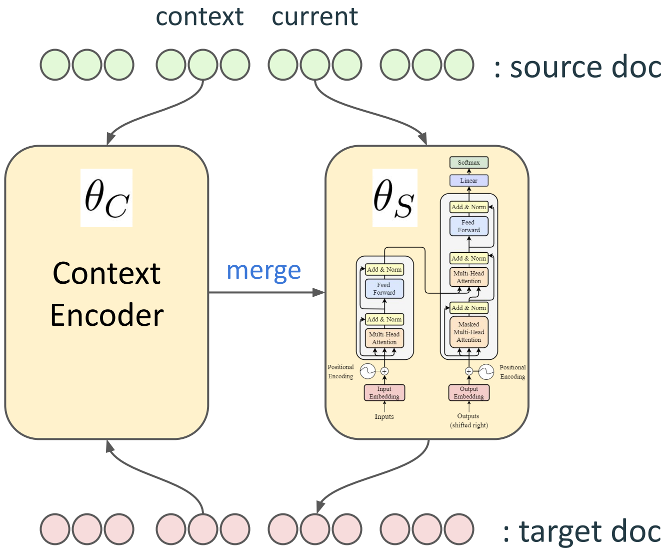


Transformer by Vaswani et al. (2017), “Attention is all you need”.

Multi-encoding approaches

contextual parameters

Trainable parameters: $\Theta = [\theta_S; \boxed{\theta_C}]$



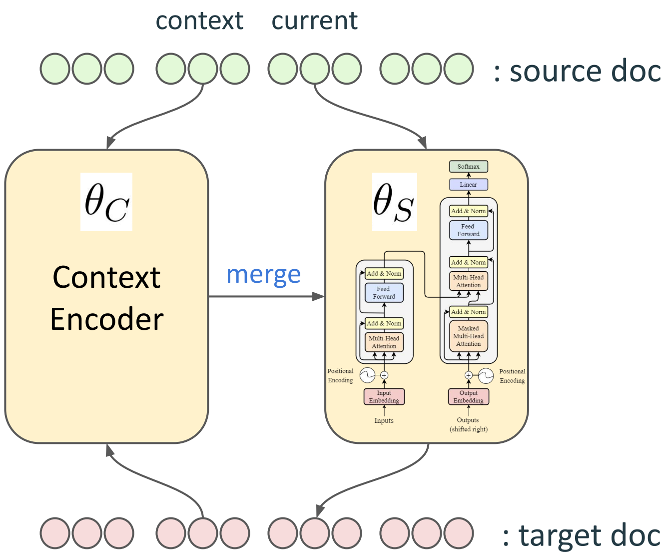
Transformer by Vaswani et al. (2017), “Attention is all you need”.

θ_S are trained on sentence-level and document-level data without θ_C;

Multi-encoding approaches

contextual parameters

Trainable parameters: $\Theta = [\theta_S; \boxed{\theta_C}]$



Transformer by Vaswani et al. (2017), "Attention is all you need".

θ_S are trained on sentence-level and document-level data without θ_C;

θ_C are trained on document-level data while θ_S are freezed;

Multi-encoding approaches

Strengths	Weaknesses
Efficient generation and processing with self-attention.	More parameters.
Self-attention is not *distracted* by context [Bao et al., 2021]: it can focus on intra-sentential linguistic relationships, which are the most important.	Kim et al. (2019), Li et al. (2020) and Lopes et al. (2020) found multi-encoding approaches to underperform context-agnostic NMT.

Multi-encoding approaches

Strengths	Weaknesses
Efficient generation and processing with self-attention.	More parameters.
Self-attention is not *distracted* by context [Bao et al., 2021]: it can focus on intra-sentential linguistic relationships, which are the most important.	Kim et al. (2019), Li et al. (2020) and Lopes et al. (2020) found multi-encoding approaches to underperform context-agnostic NMT.

Multi-encoding approaches

Strengths	Weaknesses
Efficient generation and processing with self-attention.	More parameters.
Self-attention is not *distracted* by context [Bao et al., 2021]: it can focus on intra-sentential linguistic relationships, which are the most important.	Kim et al. (2019), Li et al. (2020) and Lopes et al. (2020) found multi-encoding approaches to underperform context-agnostic NMT.

Contextual parameters are hard to train !

Double challenge of sparsity

Strengths	Weaknesses
Efficient generation and processing with self-attention.	More parameters.
Self-attention is not *distracted* by context [Bao et al., 2021]: it can focus on intra-sentential linguistic relationships, which are the most important.	**Contextual parameters are hard to train !**

Double challenge of sparsity

1. The **sparsity of the training signal**: words needing context to be correctly translated are sparse;
   - most of the words of a sentence can be translated without context → **scarce training signal**.

Double challenge of sparsity

1. The **sparsity of the training signal**: words needing context to be correctly translated are sparse;
    - most of the words of a sentence can be translated without context → **scarce training signal**.
2. the **sparsity of context words that are useful** for contextualization
    - most of the context is useless
      → distracting the model from retrieving useful information → **training is hard**.

Double challenge of sparsity

1. The **sparsity of the training signal**: words needing context to be correctly translated are sparse;
   - most of the words of a sentence can be translated without context → **scarce training signal**.
2. the **sparsity of context words that are useful** for contextualization
   - most of the context is useless
     → distracting the model from retrieving useful information → **training is hard**.

Trivial solution: more data?

Double challenge of sparsity

1. The **sparsity of the training signal**: words needing context to be correctly translated are sparse;
    - most of the words of a sentence can be translated without context ⟶ **scarce training signal**.
2. the **sparsity of context words that are useful** for contextualization
    - most of the context is useless
      ⟶ distracting the model from retrieving useful information ⟶ **training is hard**.

Trivial solution: more data?

However:

- Document-level parallel data are scarcely available.
- **Inefficient** because of the double challenge of sparsity.

Proposed approach: Divide and Rule

We propose a solution that addresses the double challenge of sparsity **directly**:

Proposed approach: Divide and Rule

We propose a solution that addresses the double challenge of sparsity **directly**:

X_j Good morning Mr. President , how are you today ?

Y_j *Bonjour Monsieur le Président , comment allez-vous aujourd' hui ?*

$\forall (x_j, y_j) \in C_{train}$

Proposed approach: Divide and Rule

We propose a solution that addresses the double challenge of sparsity **directly**:

X_j Good morning Mr. President , how are you today ?

Y_j *Bonjour Monsieur le Président , comment allez-vous aujourd' hui ?*

↓

$X_{j,1}$ Good morning Mr. President , **\n** how are you today ? $X_{j,2}$

$Y_{j,1}$ *Bonjour Monsieur le Président , **\n** comment allez-vous aujourd' hui ?* $Y_{j,2}$

$\forall (X_j, Y_j) \in C_{train}$

Proposed approach: Divide and Rule

We propose a solution that addresses the double challenge of sparsity **directly**:

X_j Good morning <u>Mr. President</u> , how <u>are you</u> today ?

Y_j *Bonjour <u>Monsieur le Président</u> , comment <u>allez-vous</u> aujourd' hui ?*

$X_{j,1}$ Good morning <u>Mr. President</u> , **\n** how <u>are you</u> today ? $X_{j,2}$

$Y_{j,1}$ *Bonjour <u>Monsieur le Président</u> ,* **\n** *comment <u>allez-vous</u> aujourd' hui ?* $Y_{j,2}$

1) Pre-train on split data. $\forall (x_j, y_j) \in C_{train}$

Proposed approach: Divide and Rule

We propose a solution that addresses the double challenge of sparsity **directly**:

2) Train on original data.

X_j

Good morning Mr. President , how are you today ?

Y_j *Bonjour Monsieur le Président , comment allez-vous aujourd' hui ?*

$X_{j,1}$ $X_{j,2}$

Good morning Mr. President , **\n** how are you today ?

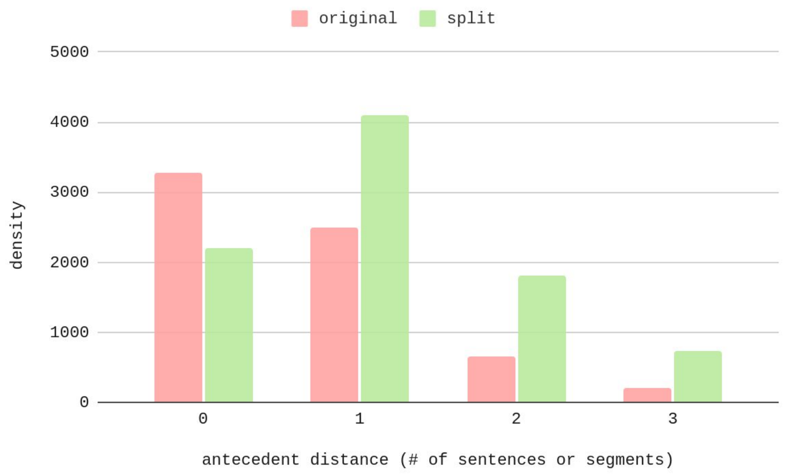
$Y_{j,1}$ *Bonjour Monsieur le Président , **\n** comment allez-vous aujourd' hui ?* $Y_{j,2}$

1) Pre-train on split data.

$\forall (x_j, y_j) \in C_{train}$

Proof of concept

How does the distribution of pronominal antecedents change when sentences are split in a half?



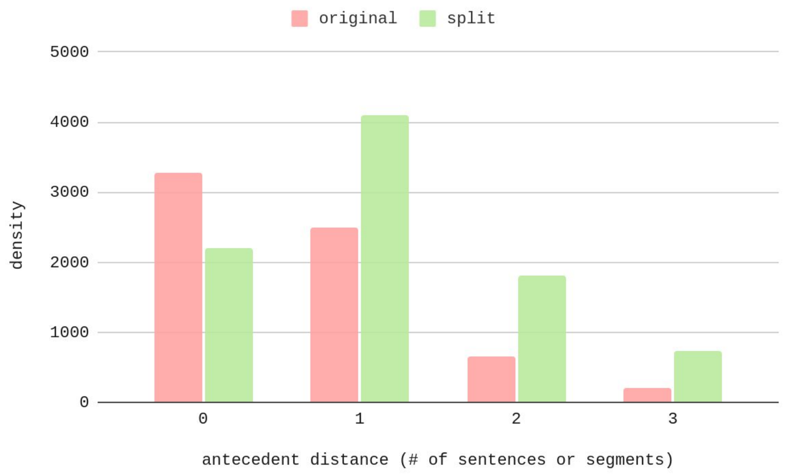
Density of pronominal antecedents by distance; Opensubs18. *Density = occurrences / # tokens to attend*.

Proof of concept

How does the distribution of pronominal antecedents change when sentences are split in a half?

1. **More cases of context-dependent anaphoric pronouns** because training sequences become incomplete segments:

   ➡ reduced sparsity of the training signal.



Density of pronominal antecedents by distance; Opensubs18. *Density = occurrences / # tokens to attend*.

Proof of concept

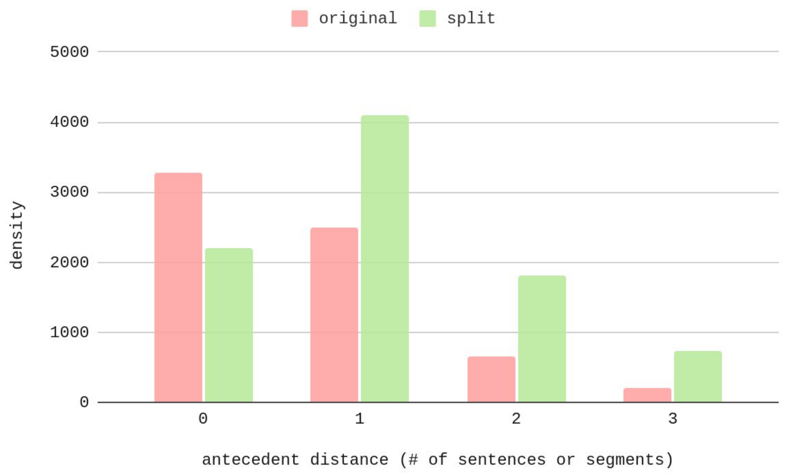
How does the distribution of pronominal antecedents change when sentences are split in a half?

1. **More cases of context-dependent anaphoric pronouns** because training sequences become incomplete segments:

   ➜ reduced sparsity of the training signal.

2. **Denser cases of pronominal antecedents** because training sequences become shorter:

   ➜ reduced sparsity of relevant context.



Density of pronominal antecedents by distance; Opensubs18. *Density = occurrences / # tokens to attend*.

Experimental Setup

Models

base: Transformer-base with parameters θ_S ;

K1: current sentence + 1 past **source context** sentences;

K3: current sentence + 3 past **source context** sentences.

Experimental Setup

Models

base: Transformer-base with parameters θ_S ;

K1: current sentence + 1 past **source context** sentences;

K3: current sentence + 3 past **source context** sentences.

Multi-encoding architecture by Miculicich et al. (2018).

Total parameters: $\Theta = [\theta_S; \theta_C]$

Experimental Setup

Models

base: Transformer-base with parameters θ_S ;

K1: current sentence + 1 past **source context** sentences;

K3: current sentence + 3 past **source context** sentences.

Multi-encoding architecture by Miculicich et al. (2018).

Total parameters: $\Theta = [\theta_S; \theta_C]$

Data

θ_S are trained on sentence-level + document-level data;

θ_C are trained on document-level data while θ_S are freezed:

Experimental Setup

Models

base: Transformer-base with parameters θ_S ;

K1: current sentence + 1 past **source context** sentences;

K3: current sentence + 3 past **source context** sentences.

Multi-encoding architecture by Miculicich et al. (2018).

Total parameters: $\Theta = [\theta_S; \theta_C]$

Data

θ_S are trained on sentence-level + document-level data;

θ_C are trained on document-level data while θ_S are freezed:

→ "Lower" resource setting (0.2-0.6M sents);

→ "Higher" resource setting (2-6M sents).

Experimental Setup

Models

base: Transformer-base with parameters θ_S ;

K1: current sentence + 1 past **source context** sentences;

K3: current sentence + 3 past **source context** sentences.

Multi-encoding architecture by Miculicich et al. (2018).

Total parameters: $\Theta = [\theta_S; \theta_C]$

Data

θ_S are trained on sentence-level + document-level data;

θ_C are trained on document-level data while θ_S are freezed:

➔ "Lower" resource setting (0.2-0.6M sents);

➔ "Higher" resource setting (2-6M sents).

3 language pairs: English $\longrightarrow$ Russian/German/French.

Experimental Setup

Evaluation

BLEU on test set.

[Papinei et al., 2020]

Experimental Setup

Evaluation

BLEU on test set.

[Papinei et al., 2020]

BLEU is ill-equipped for measuring context-aware translation improvements, which affect a few words only -> targeted evaluation is necessary to appreciate model differences.

Experimental Setup

Evaluation

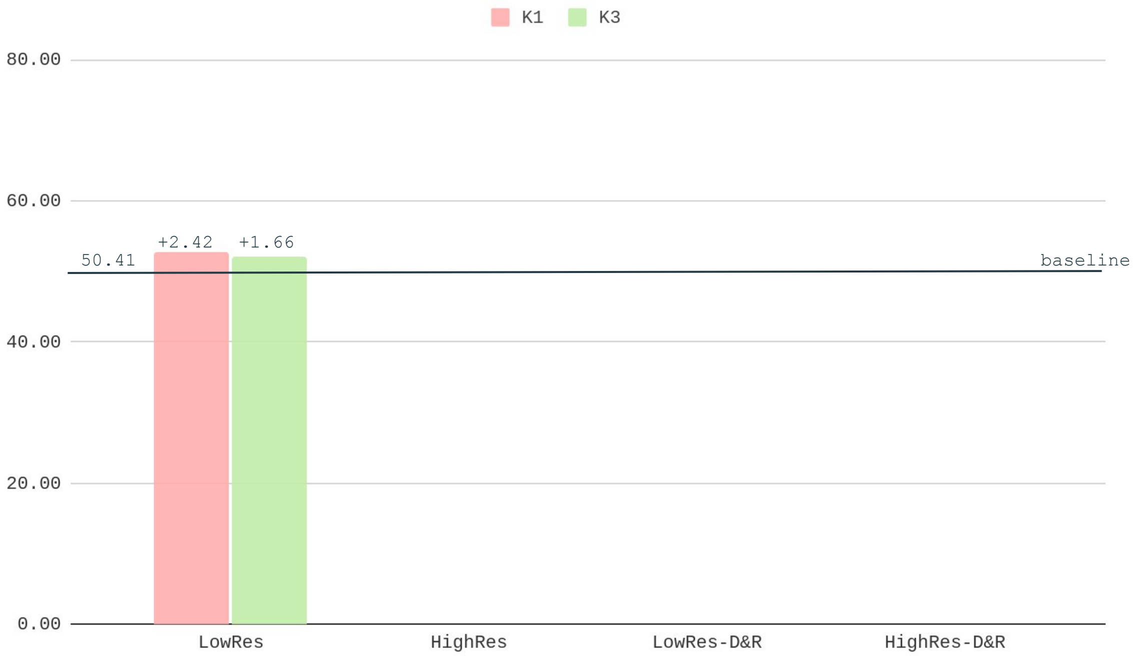
BLEU on test set.

[Papinei et al., 2020]

BLEU is ill-equipped for measuring context-aware translation improvements, which affect a few words only -> targeted evaluation is necessary to appreciate model differences.

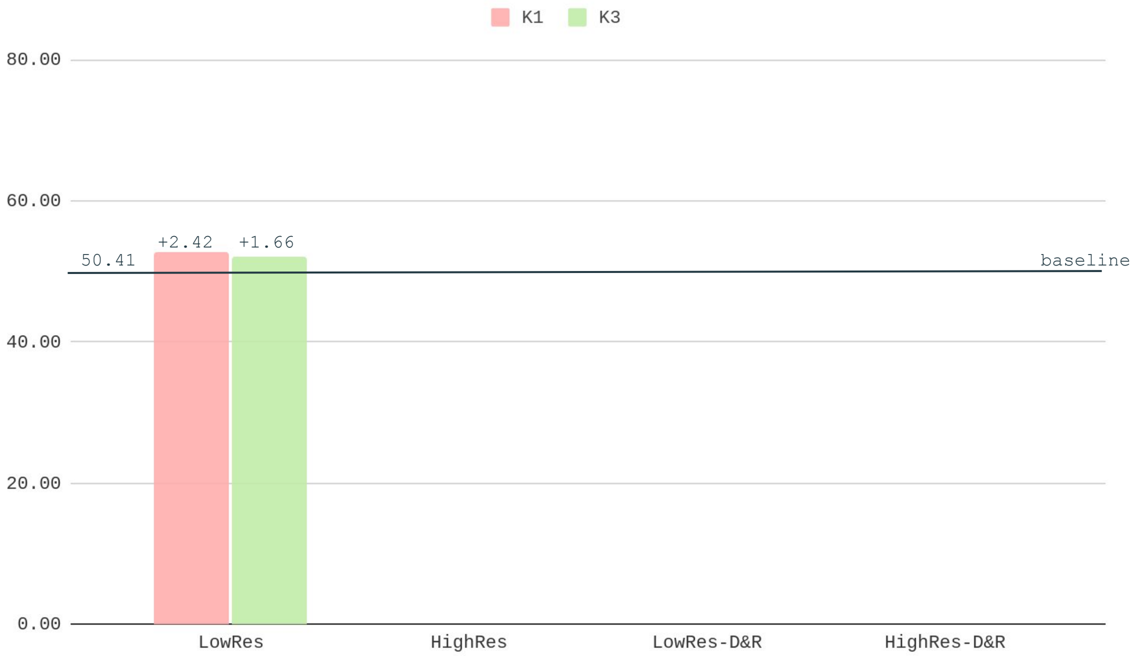
\+ Accuracy on **contrastive test sets** for the evaluation of discourse phenomena disambiguation.

- **ContraPro** (En-De/Fr)**:** anaphoric pronouns [Muller et al., 2018; Lopes et al., 2020].
- **Voita** (En-Ru)**:** verb-phrase ellipsis [Voita et al., 2019].



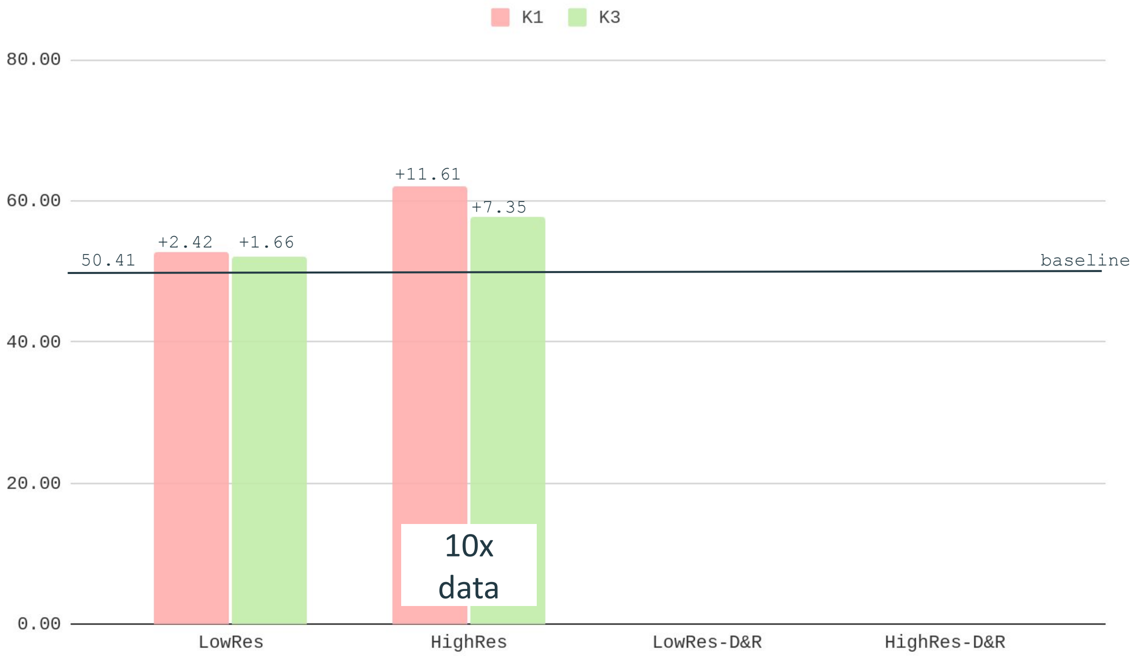
Accuracy on contrastive test sets for the translation of discourse phenomena, averaged across the three language pairs En ⟶ Ru/De/Fr

Low Res is not enough



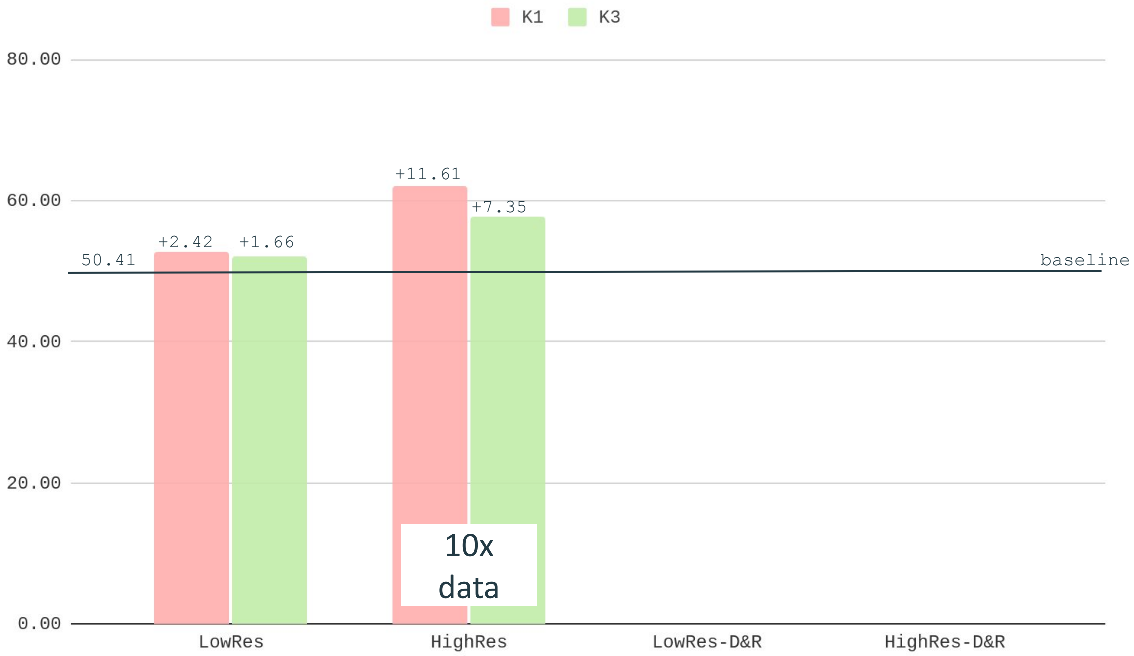
Accuracy on contrastive test sets for the translation of discourse phenomena, averaged across the three language pairs En ⟶ Ru/De/Fr

High Res is a solution



Accuracy on contrastive test sets for the translation of discourse phenomena, averaged across the three language pairs En ⟶ Ru/De/Fr

High Res is a solution

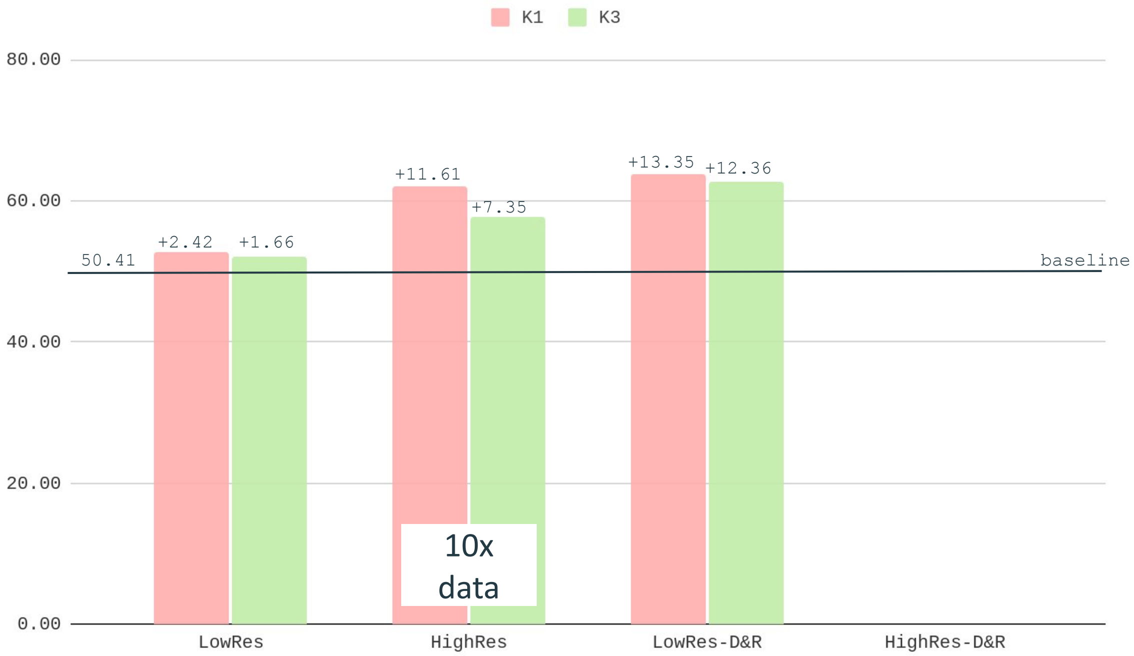


Accuracy on contrastive test sets for the translation of discourse phenomena, averaged across the three language pairs En ⟶ Ru/De/Fr

Many works in the literature trained and compared multi-encoding models on IWSLT.

➔ More training is needed

Divide and Rule is an **efficient solution**

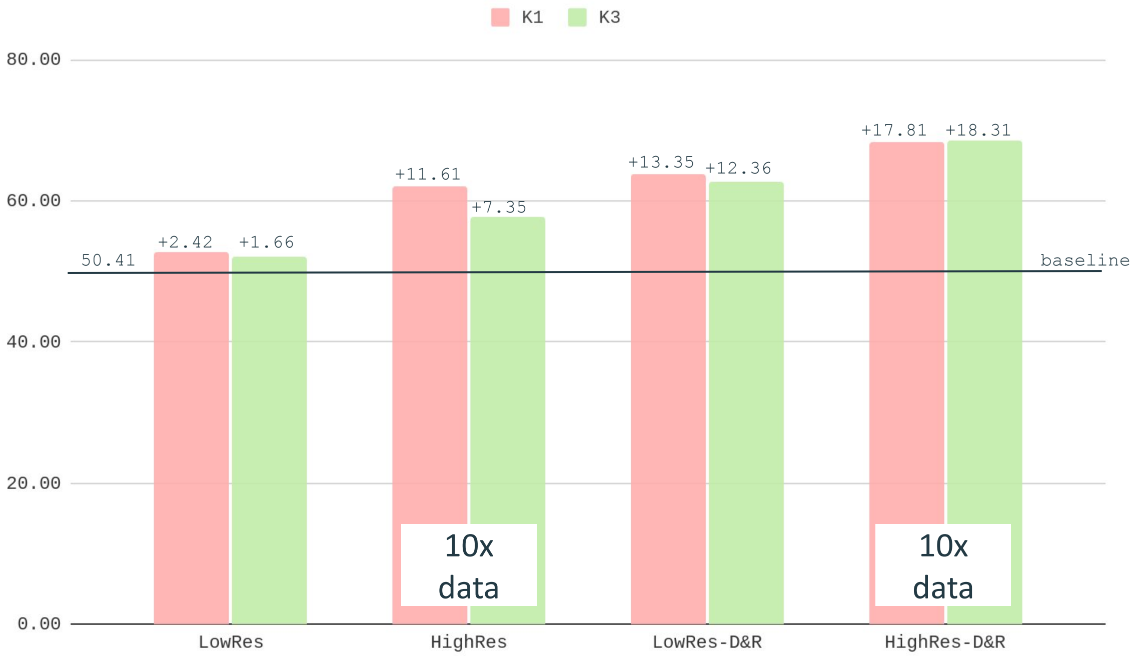


Accuracy on contrastive test sets for the translation of discourse phenomena, averaged across the three language pairs En ⟶ Ru/De/Fr

Many works in the literature trained and compared multi-encoding models on IWSLT.

➔ More training is needed

Divide and Rule is an **effective solution**

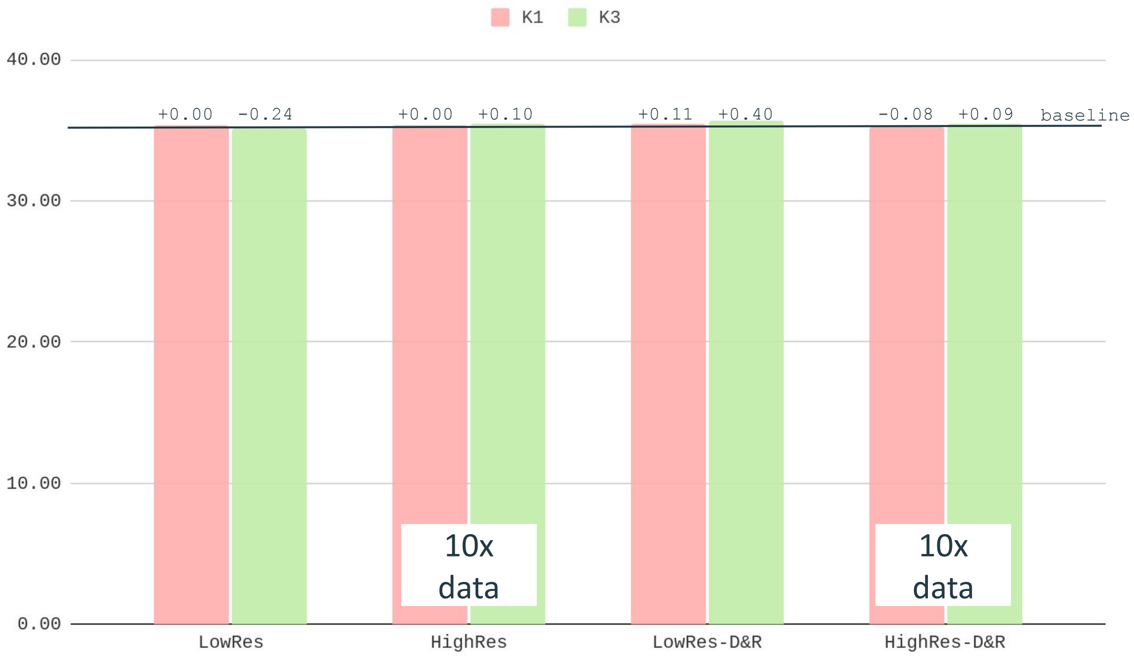


Accuracy on contrastive test sets for the translation of discourse phenomena, averaged across the three language pairs En ⟶ Ru/De/Fr

Many works in the literature trained and compared multi-encoding models on IWSLT.

➔ More training is needed

Divide and Rule is an **effective solution**



BLEU on the test sets, averaged across the three language pairs En ⟶ Ru/De/Fr

Many works in the literature trained and compared multi-encoding models on IWSLT.

➔ More training is needed

BLEU is virtually constant across the training settings.

➔ Average translation quality is constant while the modeling of inter-sentential discourse phenomena is improving.

Where to split?

Middle

Good morning Mr. President , how are you today ?

Bonjour Monsieur le Président , comment allez-vous aujourd' hui ?

1 2 3 4 5 1 2 3 4 5

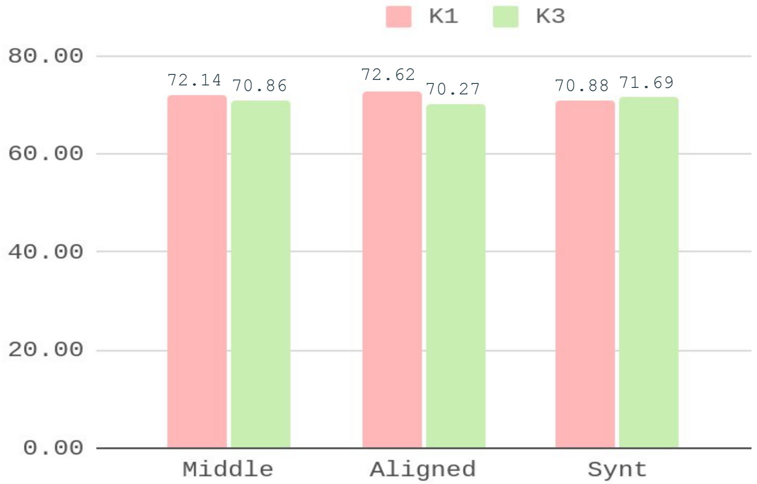
Good morning Mr. President , **\n** how are you today ?

1 2 3 4 5 1 2 3 4 5

Bonjour Monsieur le Président , **\n** *comment allez-vous aujourd' hui ?*

Where to split?

Middle **Aligned** **Synt**

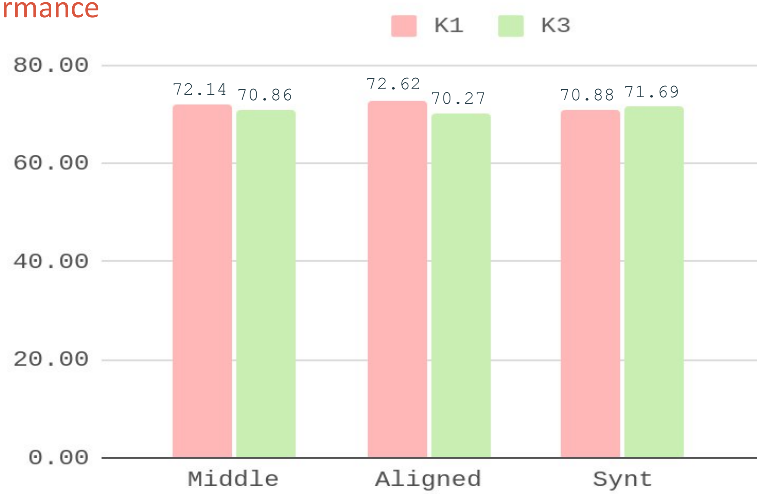


Accuracy on targeted test sets for the translation of coreferential pronouns, averaged across En → De/Fr language pairs

Where to split?

Middle **Aligned** **Synt**

Best tradeoff simplicity-performance

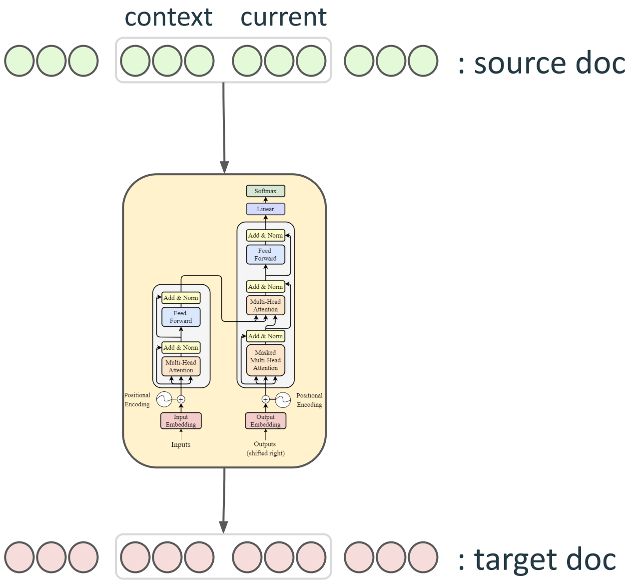


Accuracy on targeted test sets for the translation of coreferential pronouns, averaged across En → De/Fr language pairs

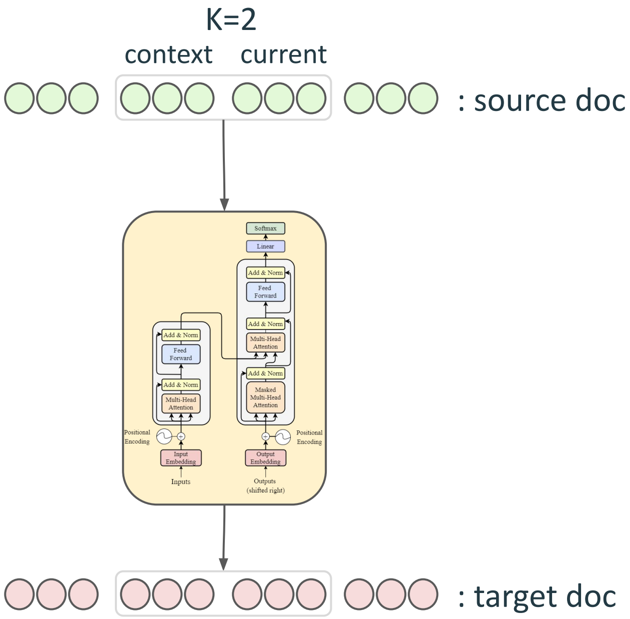
Outline

1. Introduction
2. Multi-encoding approaches
   a. Lupo, L., Dinarelli, M. and Besacier, L., **Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder NMT**, ACL 2022.
3. Concatenation approaches
   a. Lupo, L., Dinarelli, M. and Besacier, L., **Focused Concatenation for Context-Aware NMT**, WMT 2022.
   b. Lupo, L., Dinarelli, M. and Besacier, L., **Encoding Sentence Position in Context-Aware NMT with Concatenation**, Insights 2023.
4. Conclusions

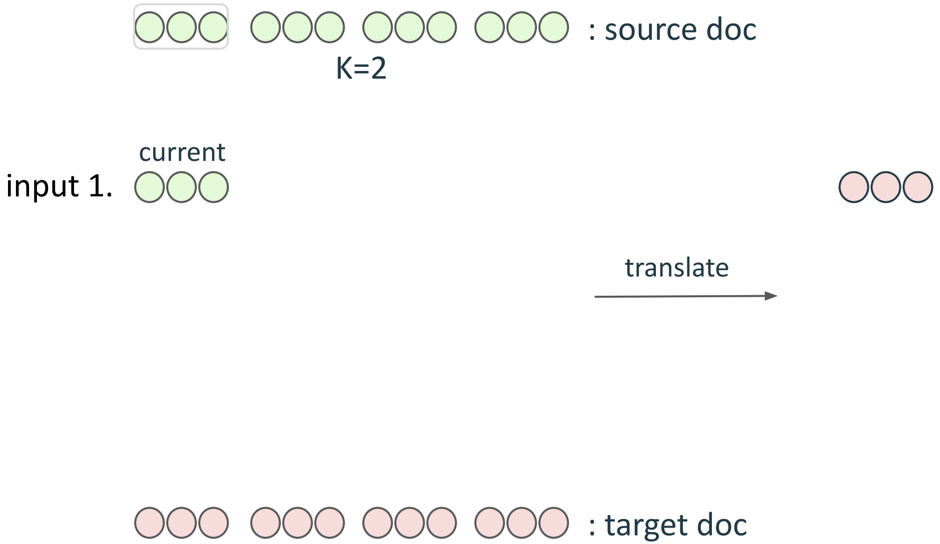
Concatenation approaches



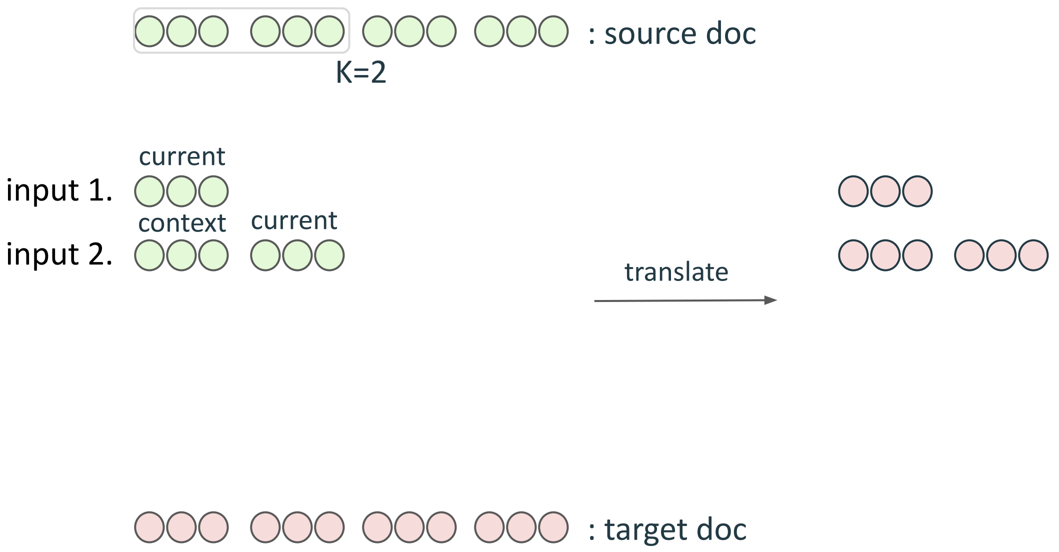
Concatenation approaches: SlidingKtoK



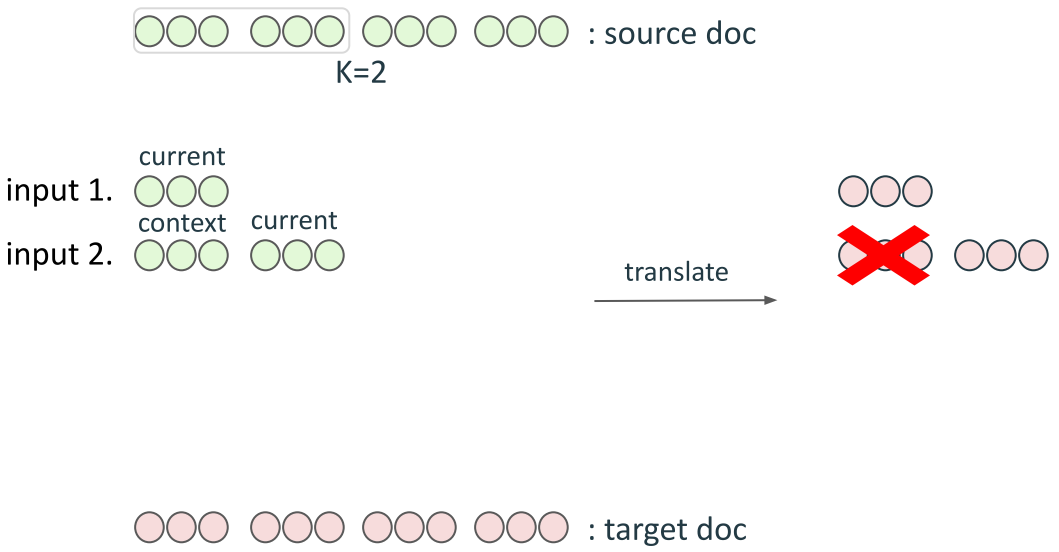
Concatenation approaches: SlidingKtoK



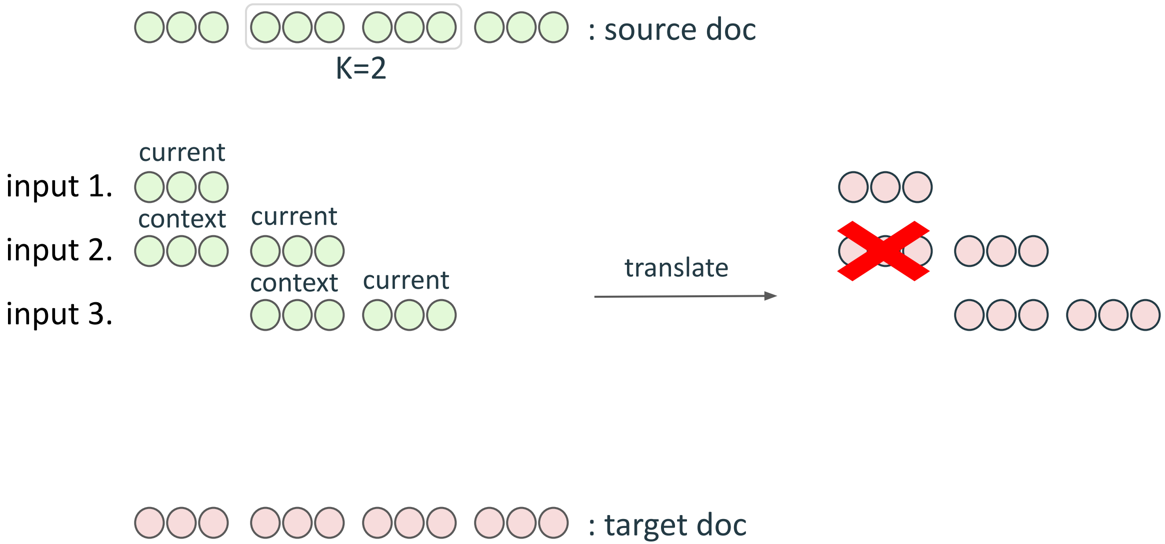
Concatenation approaches: SlidingKtoK



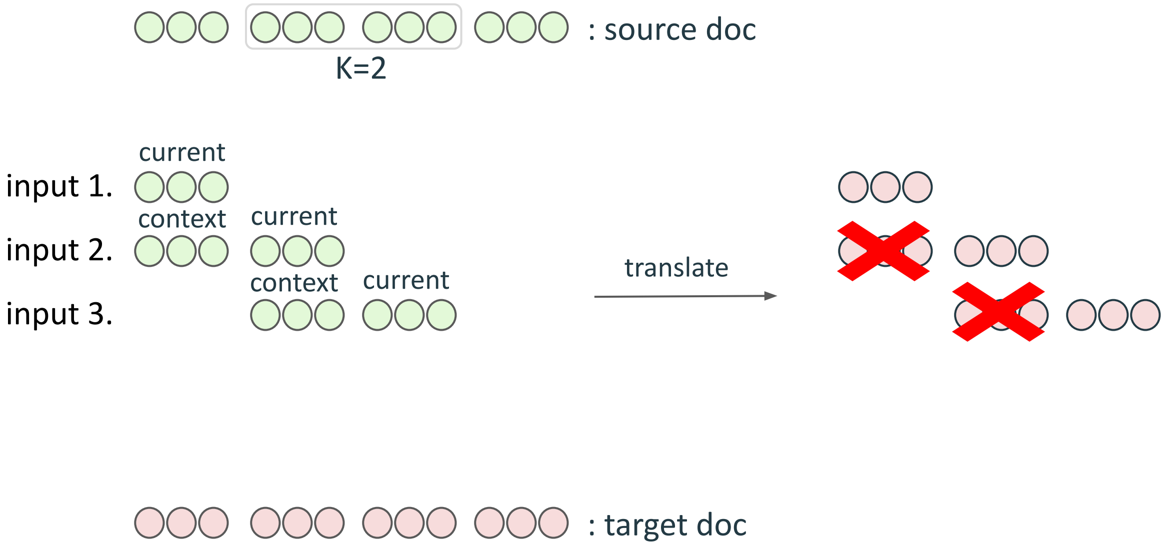
Concatenation approaches: SlidingKtoK



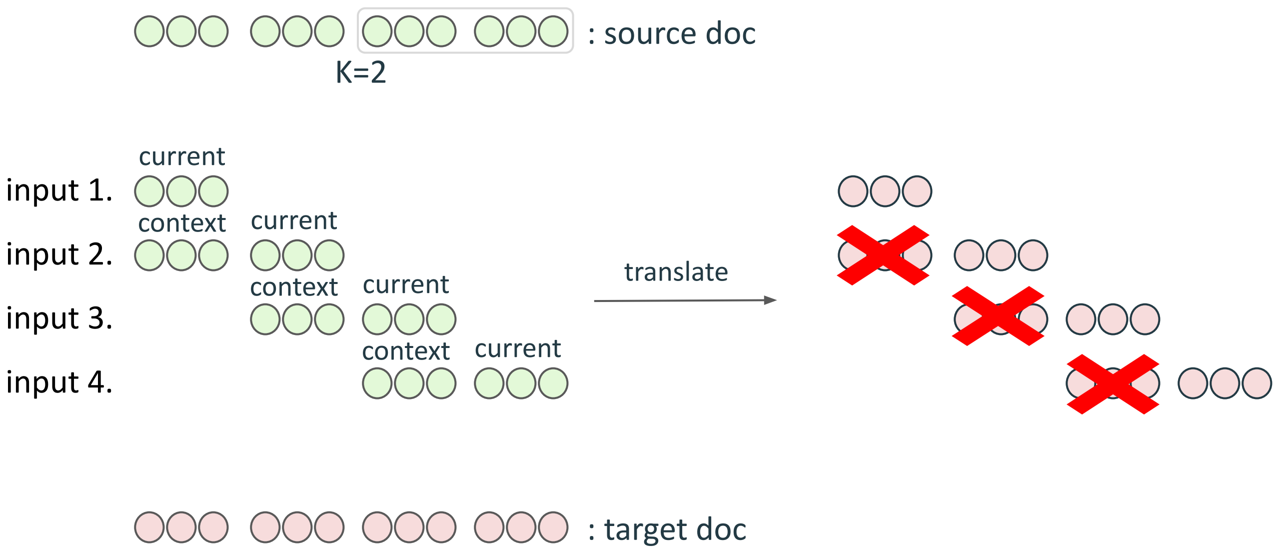
Concatenation approaches: SlidingKtoK



Concatenation approaches: SlidingKtoK



Concatenation approaches: SlidingKtoK



Concatenation approaches: SlidingKtoK

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

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

Concatenation approaches: SlidingKtoK

Strengths	Weaknesses
No extra learnable parameters added to the standard Transformer architecture.	**Attention can be *distracted* by context** instead of focusing on local relationships between tokens, which are the most important [Bao et al., 2021].
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**.

Concatenation approaches: SlidingKtoK

Strengths	Weaknesses
No extra learnable parameters added to the standard Transformer architecture.	**Attention can be *distracted* by context** instead of focusing on intra-sentential linguistic relationships, which are the most important [Bao et al., 2021].
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**.

Concatenation approaches: remedies

1. **Context discounting** in the training objective.
2. **Encoding sentence position** into token representations.

Remedy 1: context discounting

1. **Context discounting** in the training objective.
2. **Encoding sentence position** into token representations.

Remedy 1: context discounting

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

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

Remedy 1: context discounting

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

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

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

Concatenation approaches: remedies

1. **Context discounting** in the training objective.
2. **Encoding sentence position** into token representations.

Remedy 2: encoding sentence position

1. **Context discounting** in the training objective.
2. **Encoding sentence position** into token representations.
    a. Segment-shifted position embeddings.

Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Position Embedding	PE_1	PE_2	PE_3	PE_4	PE_5	PE_6	PE_7

Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Segment-shifted Position Embedding	PE_1	PE_2	PE_3	PE_{4+s}	PE_{5+s}	PE_{6+s}	PE_{7+s}

s = shift

Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Segment-shifted Position Embedding	PE_1	PE_2	PE_3	PE_{14}	PE_{15}	PE_{16}	PE_{17}

s = shift = 10

Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Segment-shifted Position Embedding	PE_1	PE_2	PE_3	PE_{14}	PE_{15}	PE_{16}	PE_{17}

s = shift = 10

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

Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Segment-shifted Position Embedding	PE_1	PE_2	PE_3	PE_{14}	PE_{15}	PE_{16}	PE_{17}

s = shift = 10

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

Remedy 2: encoding sentence position

1. **Context discounting** training objective.
2. **Encoding sentence position** into token representations.
    a. Segment-shifted position embeddings.
    b. Segment embeddings [Devlin et al., 2019].

Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Position Embedding	PE_1	PE_2	PE_3	PE_4	PE_5	PE_6	PE_7
	+	+	+	+	+	+	+
Segment Embedding	SE_2	SE_2	SE_2	SE_1	SE_1	SE_1	SE_1

Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Position Embedding	PE_1	PE_2	PE_3	PE_4	PE_5	PE_6	PE_7
	+	+	+	+	+	+	+
Segment Embedding	SE_2	SE_2	SE_2	SE_1	SE_1	SE_1	SE_1

- One-hot.
- Learned [Devlin et al., 2019].
- Sinusoidal [Vaswani et al., 2017].

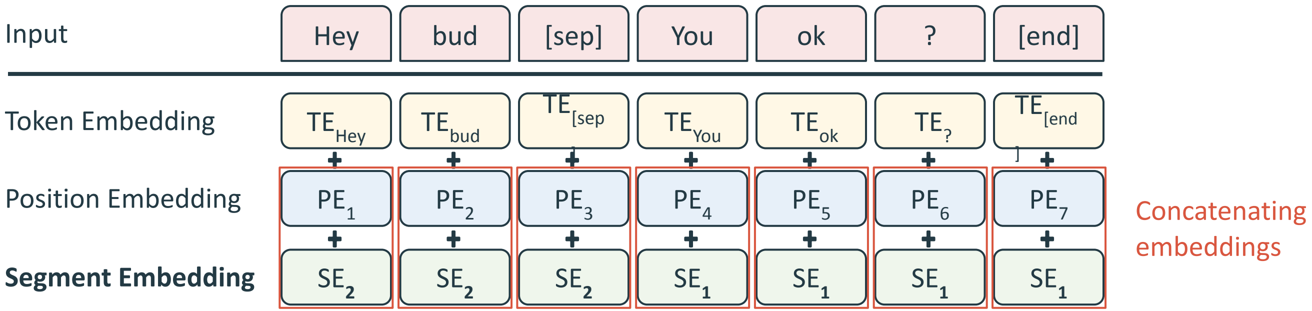
Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Position Embedding	PE_1	PE_2	PE_3	PE_4	PE_5	PE_6	PE_7
	+	+	+	+	+	+	$\boxed{+}$
Segment Embedding	SE_2	SE_2	SE_2	SE_1	SE_1	SE_1	SE_1

Can the model discern PE and SE?

- One-hot.
- Learned [Devlin et al., 2019].
- Sinusoidal [Vaswani et al., 2017].

Remedy 2: encoding sentence position



- One-hot.
- Learned [Devlin et al., 2019].
- Sinusoidal [Vaswani et al., 2017].

Remedy 2: encoding sentence position

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Position Embedding	PE_1	PE_2	PE_3	PE_4	PE_5	PE_6	PE_7
	+	+	+	+	+	+	+
Segment Embedding	SE_2	SE_2	SE_2	SE_1	SE_1	SE_1	SE_1

Concatenating embeddings requires a projection back to d_{model}

- One-hot.
- Learned [Devlin et al., 2019].
- Sinusoidal [Vaswani et al., 2017].

Remedy 2: encoding sentence position

1. **Context discounting** training objective.
2. **Encoding sentence position** into token representations.
    a. Segment-shifted position embeddings.
    b. Segment embeddings.
    c. Position-Segment Embeddings (PSE).

Remedy 2: encoding sentence position

To avoid another linear projection, we propose to reduce the dimensionality of PE and SE:

$$d_{PE} = d_{SE} = d_{model} \longrightarrow d_{PE} + d_{SE} = d_{model}$$

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Position Embedding	PE_1	PE_2	PE_3	PE_4	PE_5	PE_6	PE_7
	+	+	+	+	+	+	+
Segment Embedding	SE_2	SE_2	SE_2	SE_1	SE_1	SE_1	SE_1

Concatenating embeddings requires a projection back to d_{model}

- One-hot.
- Learned [Devlin et al., 2019].
- Sinusoidal [Vaswani et al., 2017].

Remedy 2: encoding sentence position

To avoid another linear projection, we propose to reduce the dimensionality of PE and SE:

$$d_{PE} = d_{SE} = d_{model} \longrightarrow d_{PE} + d_{SE} = d_{model}$$

Input	Hey	bud	[sep]	You	ok	?	[end]
Token Embedding	TE_{Hey}	TE_{bud}	$TE_{[sep]}$	TE_{You}	TE_{ok}	$TE_{?}$	$TE_{[end]}$
	+	+	+	+	+	+	+
Position-Segment Embedding	$PSE_{1,2}$	$PSE_{2,2}$	$PSE_{3,2}$	$PSE_{4,1}$	$PSE_{5,1}$	$PSE_{6,1}$	$PSE_{7,1}$

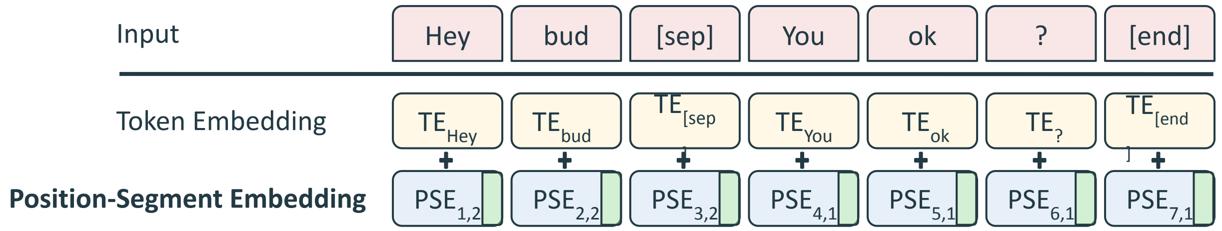
with sinusoidal PE, while SE:

- One-hot.
- Learned [Devlin et al., 2019].
- Sinusoidal [Vaswani et al., 2017].

Remedy 2: encoding sentence position

To avoid another linear projection, we propose to reduce the dimensionality of PE and SE:

$$d_{PE} = d_{SE} = d_{model} \longrightarrow d_{PE} + d_{SE} = d_{model}$$



$d_{model} = 512$

$d_{PE} = 508$
$d_{SE} = 4$

with sinusoidal PE, while SE:

- One-hot.
- Learned [Devlin et al., 2019].
- Sinusoidal [Vaswani et al., 2017].

Remedy 2: encoding sentence position

1. **Context discounting** training objective.
2. **Encoding sentence position** into token representations.
    a. Segment-shifted position embeddings.
    b. Sentence embeddings;
    c. Position-Sentence Embeddings (PSE).

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 [Voita et al., 2019]

- 6M sentence pairs from OpenSubtitles18;
- short documents of 4 sentences each.

English ⟶ German [Cettolo et al., 2012]

- 0.2M sentence pairs from IWSLT17;
- long documents of hundreds of sentences each.

Experimental Setup

Models

base: context-agnostic Transformer-base.

s4to4: sliding4to4 concatenation approach.

Data

English ⟶ Russian [Voita et al., 2019]

- 6M sentence pairs from OpenSubtitles18;
- short documents of 4 sentences each.

English ⟶ German [Cettolo et al., 2012]

- 0.2M sentence pairs from IWSLT17;
- long documents of hundreds of sentences each.

Evaluation

BLEU [Papinei et al., 2020]

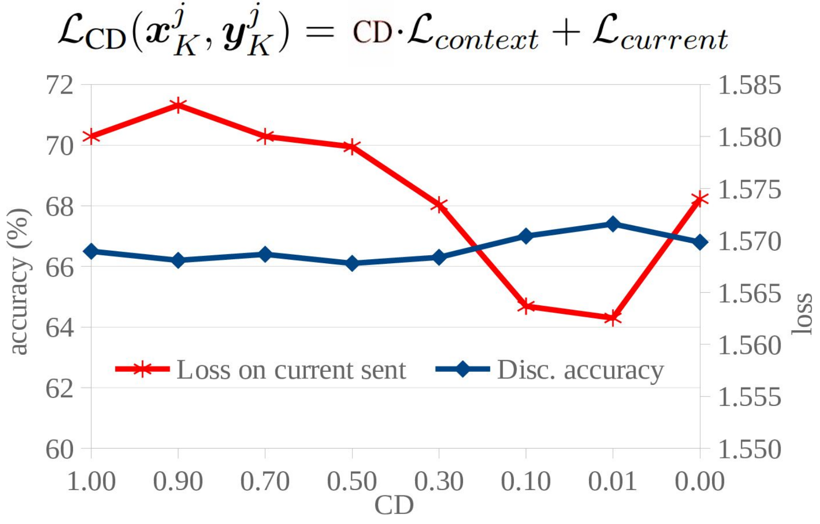
Accuracy on contrastive test sets for the disambiguation of discourse phenomena:

+ **ContraPro** (En-De): coreferential pronouns [Muller et al., 2018].
+ **Voita** (En-Ru): deixis, lexical cohesion, noun phrase ellipsis, verb-phrase ellipsis [Voita et al., 2019].

Context discounting: preliminary analysis

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

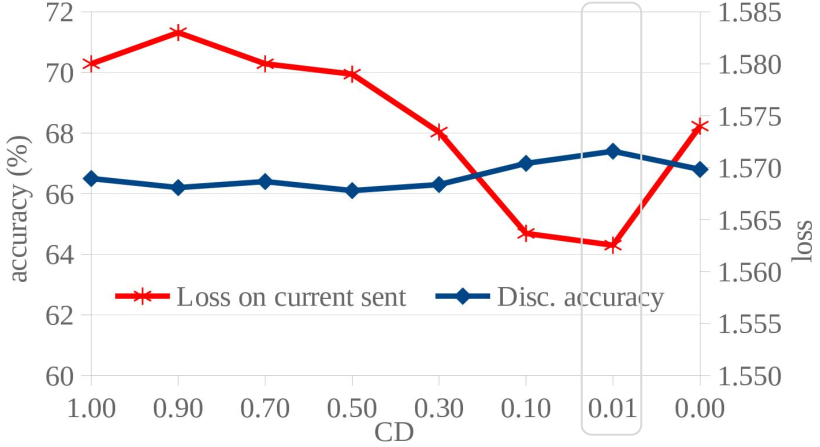
Context discounting: preliminary analysis



Evaluation of **En→Ru s4to4** trained with various levels of context discounting.

Context discounting: 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.

Context discounting: main results

	System
baselines:	base
	s4to4
	s4to4 + CD

Context discounting: main results

En→Ru	
System	BLEU
base	31.98
s4to4	32.45
s4to4 + CD	32.37
En→De	
	BLEU
base	29.63
s4to4	29.48
s4to4 + CD	29.32

baselines:

Context discounting: main results

En→Ru			
System	BLEU	Voita	
base	31.98	46.64	
s4to4	32.45	72.02	
s4to4 + CD	32.37	**73.42***	(+1.40 accuracy)
En→De			
	BLEU	ContraPro	
base	29.63	37.27	
s4to4	29.48	71.35	
s4to4 + CD	29.32	**74.31***	(+2.96 accuracy)

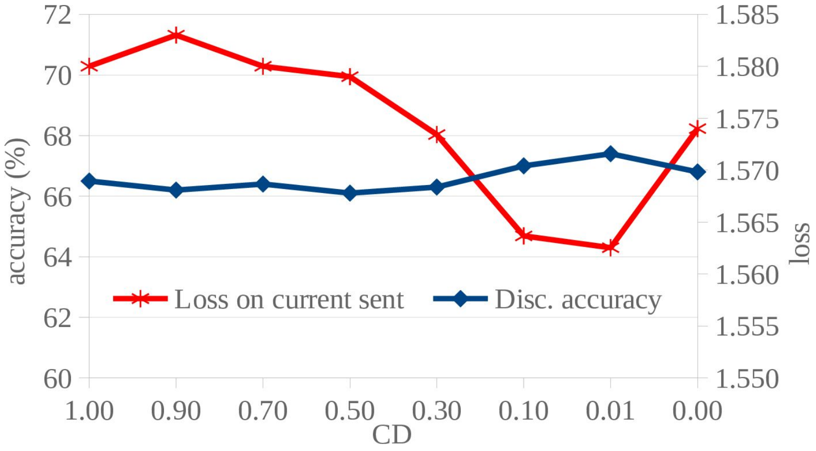
baselines: base, s4to4

Context discounting: main results

baselines:

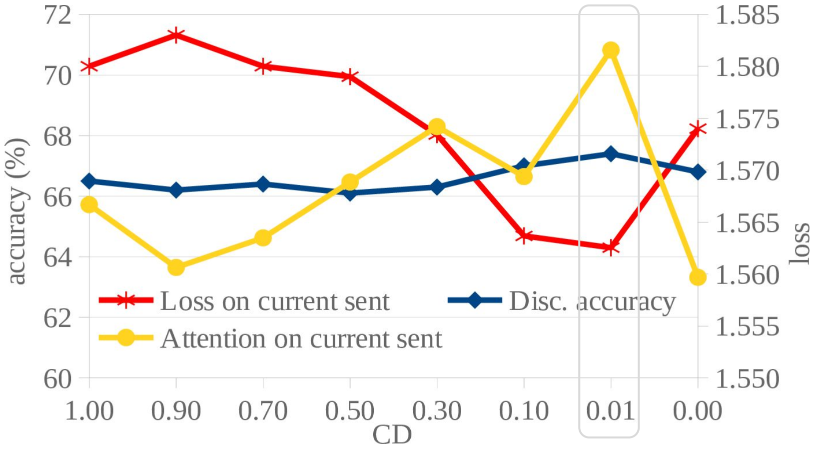
En→Ru System	BLEU	Voita	Deixis	Lex co.	Ell. inf	Ell. vp
base	31.98	46.64	50.00	45.87	51.80	27.00
s4to4	32.45	72.02	85.80	46.13	79.60	73.20
s4to4 + CD	32.37	**73.42***	**87.16***	**46.40**	**81.00**	**78.20***
En→De	**BLEU**	**ContraPro**	**d=1**	**d=2**	**d=3**	**d>3**
base	29.63	37.27	32.89	43.97	47.99	70.58
s4to4	29.48	71.35	68.89	74.96	79.58	**87.78**
s4to4 + CD	29.32	**74.31***	**72.86***	**75.96**	**80.10**	84.38

Context discounting: analysis



Evaluation of **En→Ru s4to4** trained with various levels of context discounting.

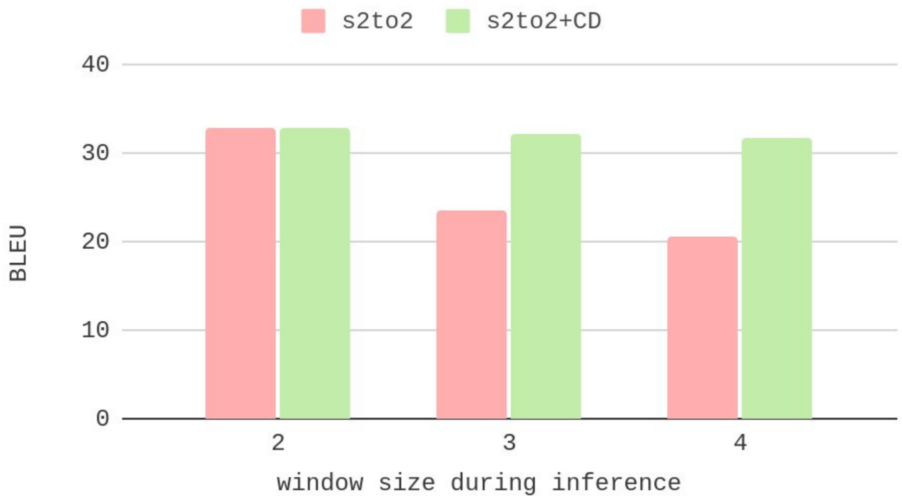
Context discounting: analysis



➔ Self-attention gets more focused.

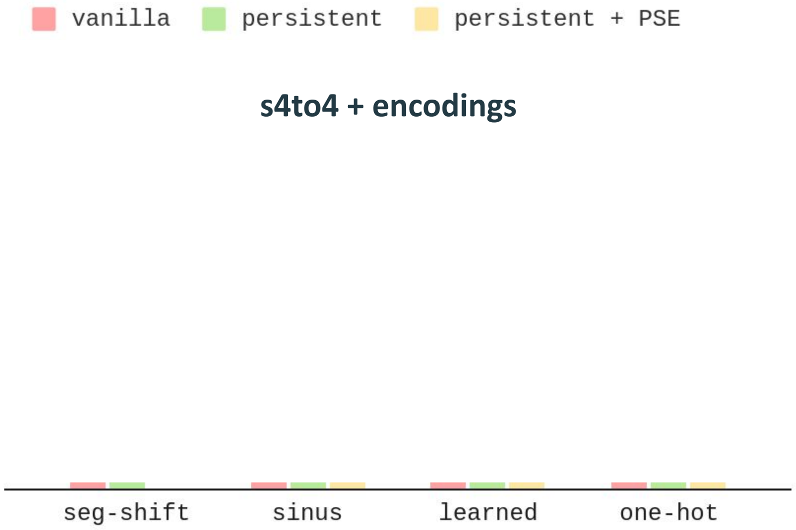
Evaluation of **En→Ru s4to4** trained with various levels of context discounting.

Context discounting: analysis

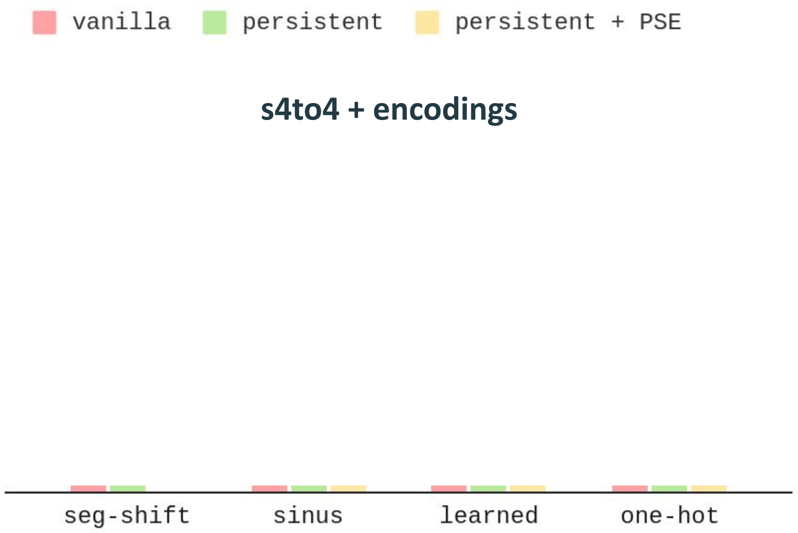


➔ Model becomes more robust to unseen context-sizes.

Encoding sentence position: main results



Encoding sentence position: main results

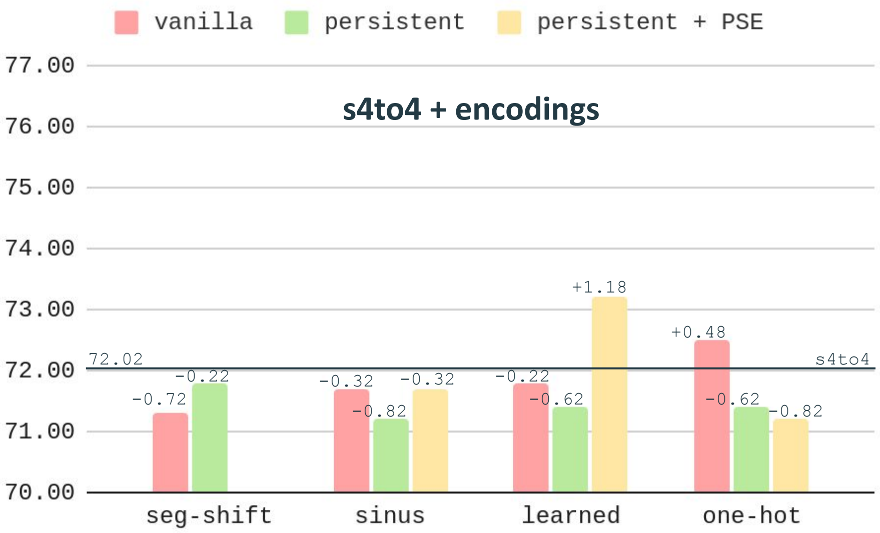


vanilla: adding encodings to the input of the 1st block

persistent: adding encodings to the input of every block

$PSE_{p,s}$ Position-Segment Embeddings

Encoding sentence position: main results



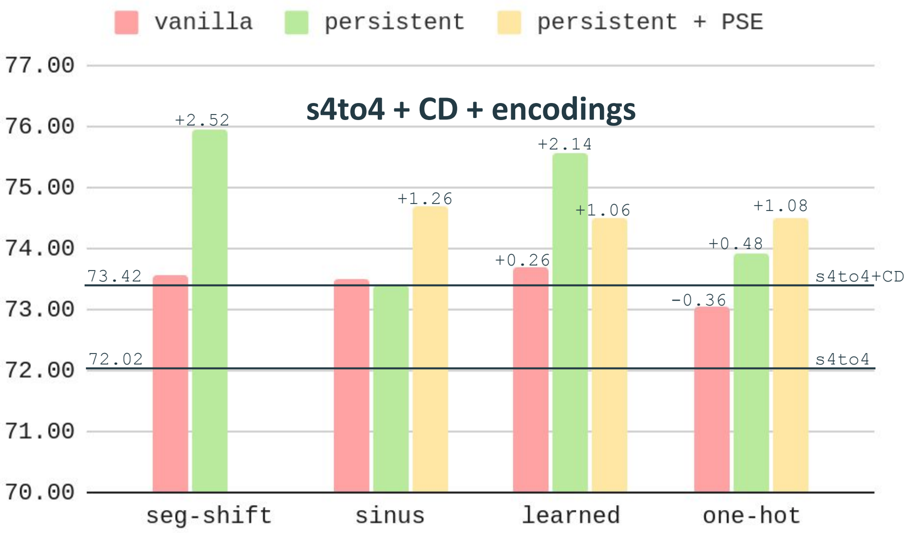
Accuracy on Voita's contrastive set on En ⟶ Ru discourse phenomena.

vanilla: adding encodings to the input of the 1st block

persistent: adding encodings to the input of every block

Position-Segment Embeddings

Encoding sentence position: main results



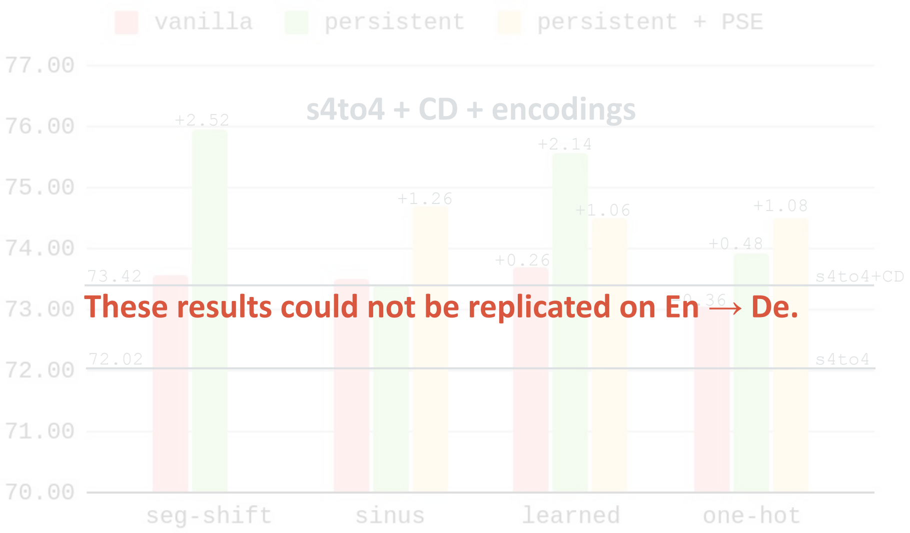
Accuracy on Voita's contrastive set on En ⟶ Ru discourse phenomena.

vanilla: adding encodings to the input of the 1st block

persistent: adding encodings to the input of every block

Position-Segment Embeddings

Encoding sentence position: main results



These results could not be replicated on En ⟶ De.

Accuracy on Voita's contrastive set on En ⟶ Ru discourse phenomena.

vanilla: adding encodings to the input of the 1st block

persistent: adding encodings to the input of every block

Position-Segment Embeddings

Benchmarking

En→Ru	
System[6]	Voita
Chen et al. (2021)	55.61
Sun et al. (2022)	58.13
Zheng et al. (2020)	63.30
Kang et al. (2020)	73.46
Zhang et al. (2020)	75.61
s4to4 + shift$_{pers}$ + CD	**75.94**

Benchmarking

En→Ru

System[6]	Voita
Chen et al. (2021)	55.61
Sun et al. (2022)	58.13
Zheng et al. (2020)	63.30
Kang et al. (2020)	73.46
Zhang et al. (2020)	75.61
s4to4 + shift$_{pers}$ + CD	**75.94**

En→De

System[6]	ContraPro
Maruf et al. (2019)	45.04
Voita et al. (2018)[7]	49.04
Stojanovski and Fraser (2019)	57.64
Müller et al. (2018)	59.51
Lupo et al. (2022a)	61.09
Lopes et al. (2020)	70.8
Majumder et al. (2022)	78.00
Fernandes et al. (2021)	80.35
Huo et al. (2020)	**82.60**
s4to4 + CD	**82.54**

Benchmarking

En→Ru

System[6]	Voita
Chen et al. (2021)	55.61
Sun et al. (2022)	58.13
Zheng et al. (2020)	63.30
Kang et al. (2020)	73.46
Zhang et al. (2020)	75.61
s4to4 + shift$_{pers}$ + CD	**75.94**

En→De

System[6]	ContraPro
Maruf et al. (2019)	45.04
Voita et al. (2018)[7]	49.04
Stojanovski and Fraser (2019)	57.64
Müller et al. (2018)	59.51
Lupo et al. (2022a)	61.09
Lopes et al. (2020)	70.8
Majumder et al. (2022)	78.00
Fernandes et al. (2021)	80.35
Huo et al. (2020)	**82.60**
s4to4 + CD	**82.54**

> x 10
training data

Outline

1. Introduction
2. Multi-encoding approaches
   a. Lupo, L., Dinarelli, M. and Besacier, L., **Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder NMT**, ACL 2022.
3. Concatenation approaches
   a. Lupo, L., Dinarelli, M. and Besacier, L., **Focused Concatenation for Context-Aware NMT**, WMT 2022.
   b. Lupo, L., Dinarelli, M. and Besacier, L., **Encoding Sentence Position in Context-Aware NMT with Concatenation**, Insights 2023.
4. $\boxed{\text{Conclusions}}$

Contributions

1. **Identified challenges** in both multi-encoding and concatenation approaches;

Contributions

1. **Identified challenges** in both multi-encoding and concatenation approaches;
2. **Proposed and evaluated remedies**, exploring different aspects of these approaches:

Contributions

1. **Identified challenges** in both multi-encoding and concatenation approaches;
2. **Proposed and evaluated remedies**, exploring different aspects of these approaches:
   a. the **training data** - Divide and Rule for multi-encoding approaches

Contributions

1. **Identified challenges** in both multi-encoding and concatenation approaches;
2. **Proposed and evaluated remedies**, exploring different aspects of these approaches:
   a. the **training data** - Divide and Rule for multi-encoding approaches
   b. the **training objective** - Context discounting for concatenation approaches

Contributions

1. **Identified challenges** in both multi-encoding and concatenation approaches;
2. **Proposed and evaluated remedies**, exploring different aspects of these approaches:
   a. the **training data** - Divide and Rule for multi-encoding approaches
   b. the **training objective** - Context discounting for concatenation approaches
   c. the **architecture** - Sentence position encodings for concatenation approaches;

Contributions

1. **Identified challenges** in both multi-encoding and concatenation approaches;
2. **Proposed and evaluated remedies**, exploring different aspects of these approaches:
    a. the **training data** - Divide and Rule for multi-encoding approaches
    b. the **training objective** - Context discounting for concatenation approaches
    c. the **architecture** - Sentence position encodings for concatenation approaches;
3. **Improved understanding** of context-aware NMT approaches through analysis.

Perspectives

1. **Long-range arena**: contrastive test sets for the evaluation of wider-context-aware NMT, including:
    a. long-context-dependent discourse phenomena;

Perspectives

1. **Long-range arena**: contrastive test sets for the evaluation of longer-context-aware NMT, including:
   a. long-context-dependent discourse phenomena;
2. **Large multilingual language models** (GPT3, Bloom, LLaMa) **as automatic post editors**: from context-agnostic NMT document translations to coherent translations.
   a. Prompt engineering.
   b. Inclusion of meta-data such as authors' information or a glossary for domain-specific terminology constraints.
   c. Fine-tuning on DocRepair-like training data [Voita et al., 2019b].

Thank you.

References

Bao, G., Zhang, Y., Teng, Z., Chen, B., and Luo, W. (2021). G-transformer for document-level machine translation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3442–3455, Online. Association for Computational Linguistics.

Castilho, S., Popović, M., and Way, A. (2020). On context span needed for machine translation evaluation. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 3735–3742, Marseille, France. European Language Resources Association.

Kim, Y., Tran, D. T., and Ney, H. (2019). When and why is document-level context useful in neural machine translation? In Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), pages 24–34, Hong Kong, China. Association for Computational Linguistics.

Läubli, S., Sennrich, R., and Volk, M. (2018). Has machine translation achieved human parity? a case for document-level evaluation. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Li, B., Liu, H., Wang, Z., Jiang, Y., Xiao, T., Zhu, J., Liu, T., and Li, C. (2020). Does multi-encoder help? a case study on context-aware neural machine translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3512–3518, Online. Association for Computational Linguistics.

Lopes, A., Farajian, M. A., Bawden, R., Zhang, M., and Martins, A. F. T. (2020). Document-level neural MT: A systematic comparison. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, pages 225–234, Lisboa, Portugal. European Association for Machine Translation.

Lupo, L., Dinarelli, M. and Besacier, L. (2022). Divide and Rule: Effective Pre-Training for Context-Aware Multi-Encoder Translation Models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4557-4672, Dublin, Ireland.

Lupo, L., Dinarelli, M. and Besacier, L. (2022). Focused Concatenation for Context-Aware Neural Machine Translation. In Proceedings of the Seventh Conference on Machine Translation (WMT), pages 830–842, Abu Dhabi, December 7–8, 2022. Association for Computational Linguistics.

Maruf, S., Saleh, F., and Haffari, G. (2021). A Survey on Document-level Neural Machine Translation: Methods and Evaluation. ACM Computing Surveys, 54(2):45:1–45:36.

Müller, M., Rios, A., Voita, E., and Sennrich, R. (2018). A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

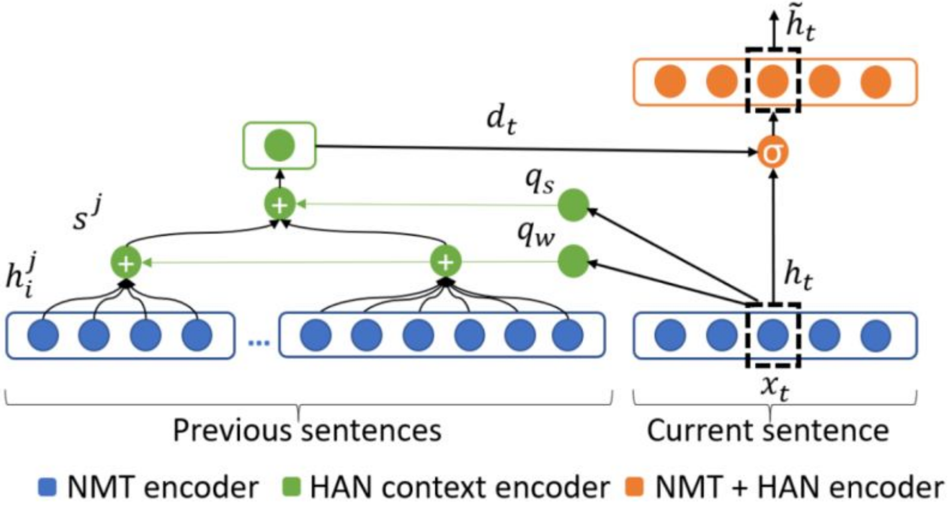
Toral, A., Castilho, S., Hu, K., and Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Voita, E., Sennrich, R., and Titov, I. (2019b). Context-aware monolingual repair for neural machine translation. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 877–886, Hong Kong, China. Association for Computational Linguistics.

HAN



$\tilde{h}_t$
d_t
q_s
s^j
q_w
h_i^j
h_t
...
x_t
Previous sentences
Current sentence
NMT encoder
HAN context encoder
NMT + HAN encoder

Contrastive test sets

Accuracy on **contrastive test sets** for the evaluation of discourse phenomena disambiguation.

Source Context

Good morning Mr President!

Target Context

Bonjour Monsieur le Président!

Source

How <u>are you</u> today?

Translation Candidates

- Comment allez-vous aujourd'hui?
- Comment vas-tu aujourd'hui?

Data

	En→Ru		En→De		En→Fr	
	Low Res	Hig Res	Low Res	Hig Res	Low Res	Hig Res
Sentence-level train	OpenSubs2018	OpenSubs2018	WMT17	WMT17	WMT14	WMT14
Context-aware train	1/10th of OpenSubs2018	OpenSubs2018	IWSLT17	News-v12 Europarl-v7 IWSLT17	IWSLT17	News-v9 Europarl-v7 IWSLT17
Fine-tuning	-	-	-	IWSLT17	-	IWSLT17
Test (BLEU)	OpenSubs2018	OpenSubs2018	IWSLT17	IWSLT17	IWSLT17	IWSLT17
Contrastive test	EllipsisVP	EllipsisVP	ContraPro	ContraPro	ContraPro	ContraPro

Contrastive test sets [voita et al., 2019a]

(a) **EN** We haven't really spoken much since your return. Tell me, what's on your mind these days?
RU Мы не разговаривали с тех пор, как вы вернулись. Скажи мне, что у тебя на уме в последнее время?
RU My ne razgovarivali s tekh por, kak vy vernulis'. Skazhi mne, chto u tebya na ume v posledneye vremya?

(b) **EN** I didn't come to Simon's for you. I did that for me.
RU Я пришла к Саймону не ради тебя. Я сделал это для себя.
RU Ya prishla k Saymonu ne radi tebya. Ya sdelal eto dlya sebya.

Figure 1: Examples of violation of (a) T-V form consistency, (b) speaker gender consistency.
In color: (a) red – V-form, blue – T-form; (b) red – feminine, blue – masculine.

(a) **EN** You call her your friend but have you been to her home ? Her work ?
RU Ты называешь её своей подругой, но ты был у неё дома? Её работа?
RU Ty nazyvayesh' yeyo svoyey podrugoy, no ty byl u neye doma? Yeyo rabota?

(b) **EN** Veronica, thank you, but you saw what happened. We all did.
RU Вероника, спасибо, но ты видела, что произошло. Мы все хотели.
RU Veronika, spasibo, no ty videla, chto proizoshlo. My vse khoteli.

Figure 2: Examples of discrepancies caused by ellipsis. (a) wrong morphological form, incorrectly marking the noun phrase as a subject. (b) correct meaning is "see", but MT produces хотели *khoteli* ("want").

(a) **EN** Not for Julia. Julia has a taste for taunting her victims.
RU Не для Джулии. Юлия умеет дразнить своих жертв.
RU Ne dlya Dzhulii. Yuliya umeyet draznit' svoikh zhertv.

(b) **EN** But that's not what I'm talking about. I'm talking about your future.
RU Но я говорю не об этом. Речь о твоём будущем.
RU No ya govoryu ne ob etom. Rech' o tvoyom budushchem.

Figure 3: Examples of lack of lexical cohesion in MT. (a) Name translation inconsistency. (b) Inconsistent translation. Using either of the highlighted translations consistently would be good.

Testing with inconsistent context

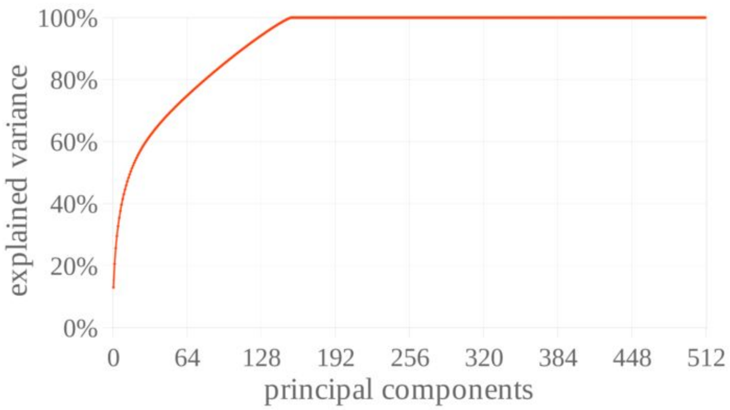
	En→De		En→Fr	
Model	BLEU	ContraPro	BLEU	ContraPro
base	32.97 (+0.00)	46.37 (0.00)	41.44 (-0.00)	79.46 (0.00)
K2	33.06 (+0.06)	46.7 (-0.35)	41.75 (-0.12)	79.05 (-0.19)
K4	32.73 (-0.13)	46.21 (-0.27)	41.47 (+0.15)	79.24 (-1.29)
K2-d&r	33.1 (-0.34)	47.6 (-12.61)	41.64 (-0.14)	78.94 (-5.12)
K4-d&r	33.05 (-0.31)	47.96 (-8.26)	41.55 (-0.13)	79.05 (-6.45)

D&R scope

- **4,000 written languages** in the world (Eberhard et al., 2021)
- Most of them can be grouped in a **few types with similar word order**, as shown by the ample literature on word order typologies (Dryer and Haspelmath, 2013; Tomlin, 2014).
- The primary order of interest is the **constituent order**, concerning the relative order of subject **(S)**, object **(O)** and verb **(V)** in a clause.
- ~40% of languages is SVO (En,Fr,Ru,De)
- ~40% of languages is SOV (De)
- ~10% of languages is VSO.

Encoding sentence position with PSE

Can we reduce the size of sinusoidal embeddings without loss of information?



Cumulative ratio of the variance explained by the principal components of the **1024 × 512** sinusoidal position embedding matrix.

$d_{model} = 512$

$d_{PE} = 508$

$d_{SE} = 4$

Context-discounting: preliminary analysis

	En→Ru			En→De	
CD	Loss	Voita^{test}	Voita^{dev}	Loss	ContraPro
1.000	1.580	69.99	66.50	1.097	70.43
0.900	1.583	70.26	66.20	1.096	69.44
0.700	1.580	70.96	66.40	1.093	70.52
0.500	1.579	70.89	66.10	1.092	70.38
0.300	1.573	71.59	66.30	1.089	72.49
0.100	1.564	71.86	67.00	**1.086**	69.58
0.010	1.563	**73.19**	67.40	1.090	**74.31**
0.009	1.563	67.30	67.30	**1.086**	71.93
0.007	**1.562**	67.90	**67.90**	1.091	72.72
0.005	**1.562**	67.00	67.00	1.110	71.25
0.003	1.563	67.20	67.20	1.105	71.13
0.001	1.563	67.50	67.50	1.104	64.53
0.000	1.574	70.34	66.80	1.191	61.14

Full context discounting?

	En→Ru					
System	Deixis	Lex co.	Ell. inf	Ell. vp	Voita	BLEU
s4to1	50.00	45.87	57.60	71.40	51.66	**32.64**
s4to4 + *CD=0*	**86.48**	**46.27**	**70.00**	**78.60**	**71.98**	28.55
	En→De					
System	d=1	d=2	d=3	d>3	ContraPro	BLEU
s4to1	36.90	46.55	49.38	69.68	40.67	**29.28**
s4to4 + *CD=0*	**57.35**	**67.81**	**71.72**	**85.29**	**61.14**	11.85

Synergies: D&R + CD

En→Ru			
System	*d&r*	Voita	BLEU
s4to4	no	72.02	32.45
s4to4 + CD	no	73.42	32.37
s4to4	yes	70.84	32.07
s4to4 + CD	yes	**74.50**	31.95

En→De			
System	*d&r*	ContraPro	BLEU
s4to4	no	71.35	29.48
s4to4 + CD	no	74.31	29.32
s4to4	yes	70.06	29.08
s4to4 + CD	yes	**74.63**	29.78

Significance testing

McNemar's test (McNemar, 1947) for comparing accuracy results on the contrastive test sets. This test is specifically designed for paired nominal observations, which is exactly the situation encountered in contrastive test sets: each system obtains a binary outcome (correct/incorrect ranking) for each contrastive example

Approximate randomization (Riezler and Maxwell, 2005) for all the other cases, e.g., for comparing BLEU scores. Approximate randomization is based on resampling and it can be applied to non-binary, non-paired scores without requiring compliance to any hypothesis about their distribution (contrarily to, for instance, the Wilcoxon test (Wilcoxon, 1946)).