Understanding and Improving Context Usage in Context-aware Translation

Graham Neubig

Carnegie Mellon University
Language Technologies Institute

Based on work with
Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary, André F.T. Martins
Why is Context Important for Translation?

We’ll have to get rid of that mole.
Why is Context Important for Translation?

*Things could start to get dangerous if the ministers find out.* We’ll have to get rid of that *mole.*
Things could start to get dangerous if the ministers find out. We’ll have to get rid of that mole.
Why is Context Important for Translation?

Could it be anything serious, Doctor? We’ll have to get rid of that mole.
Why is Context Important for Translation?

*Could it be anything serious, Doctor? We’ll have to get rid of that mole.*
Why is Context Important for Translation?

English:
Things could start to get dangerous if the ministers find out. We’ll have to get rid of that mole.

French:
Les choses pourraient commencer à devenir dangereuses si les ministres le découvraient. Nous devrons nous débarrasser de cette taupe.
Why is Context Important for Translation?

English:
Could it be anything serious, Doctor?
We’ll have to get rid of that mole.

French:
Serait-ce quelque chose de grave, docteur ?
Nous devrons nous débarrasser de cette taupe.
cet grain de beauté
Why is Context Important for Translation?

English:
So you see how bad the implications are. Yes, they are quite devastating.

French:
Vous voyez donc à quel point les implications sont mauvaises. Oui, ils sont assez dévastateurs.
So you see how bad the implications are. Yes, they are quite devastating.

Vous voyez donc à quel point les implications sont mauvaises. Oui, ils sont assez dévastateurs.
Today’s Agenda

New methods to measure context usage in MT
- Agreement with human annotations
- Conditional cross mutual information (CXMI)

New training methods to increase context usage in MT
- Supervised attention
- Contextual word dropout

→ Code & Data:
  github.com/neulab/contextual-mt
Do Context-Aware Translation Models Pay the Right Attention?

Kayo Yin, Patrick Fernandes, Danish Pruthi, Aditi Chaudhary
André F.T. Martins, Graham Neubig

(ACL 2021)
Context-Aware NMT

Source input: Have we got her report?

Yes, it’s in the infirmary already.

Context-aware NMT output: On dispose de son rapport?

Oui, elle est déjà à l’infirmerie.
Context-Aware NMT

Source input: Have we got her report? Yes, it’s in the infirmary already.

Context-aware NMT output: On dispose de son rapport? Oui, elle est déjà à l’infirmerie.
Have we got her report?
Yes, it’s in the infirmary already.

On dispose de son rapport?
Oui, elle est déjà à l’infirmérie.
Have we got her report?
Yes, it’s in the infirmary already.

On dispose de son rapport?
Oui, elle est déjà à l’infirmérie.
Outline

1. What context is useful during ambiguous translations?

2. Are models paying attention to this context or not?

3. If not, can we encourage them to do so?
Outline

1. What context is useful during translation?

2. Are models paying attention to this context or not?

3. If not, can we encourage them to do so?
User Study

Task 1 - Example 1

Source context:
Look after her a lot.
Okay.
Any questions?
Have we got her report?

Source sentence:
Yes, it’s in the infirmary already.

Target context:
Dorlotez-la.
D’accord.
Vous avez des questions?
On dispose de son rapport?

Target sentence:
Oui, ____ est à l’infirmérie.
  ○ il
  ○ elle

How confident are you?
- Not at all
- Somewhat
- Very

Source context you highlighted:
Reset Highlights

Source sentence you highlighted:
Reset Highlights

Target context you highlighted:
Reset Highlights

Target sentence you highlighted:
Reset Highlights

Mismatch between source and target side
User Study

Task 1 - Example 1

Source context:
- Look after her a lot.
- Okay.
- Any questions?
- Have we got her report?

Source sentence:
- Yes, it’s in the infirmary already.

Target context:
- Dorlottez-la.
- D'accord.
- Vous avez des questions?
- On dispose de son rapport?

Target sentence:
- Oui, ___ est à l'infirmérie.
  - il
  - elle

How confident are you?
- Not at all
- Somewhat
- Very

Source context you highlighted:
- Reset Highlights

Source sentence you highlighted:
- Reset Highlights

Target context you highlighted:
- Reset Highlights

Target sentence you highlighted:
- Reset Highlights

Mismatch between source and target side
User Study

Task 1 - Example 1

Source context:
  Look after her a lot.
  Okay.
  Any questions?
  Have we got her report?

Source sentence:
  Yes, it’s in the infirmary already.

Target context:
  Dorlotez-la.
  D’accord.
  Vous avez des questions?
  On dispose de son rapport?

Target sentence:
  Oui, ____ est à l’infirmérie.
  • il
  • elle

How confident are you?

Source context you highlighted:
Reset Highlights

Target context you highlighted:
  • rapport
Reset Highlights

Source sentence you highlighted:
Reset Highlights

Mismatch between source and target side
User Study

Task 1 - Example 33

Source context:

Source sentence: 
*Ace* of diamonds.

Target context:

Target sentence: 
As de _____
- *carreau.*
- *diamant.*

How confident are you?

- Not at all
- Somewhat
- Very

Source context you highlighted:
- Reset Highlights

Source sentence you highlighted:
  - Ace
- Reset Highlights

Target context you highlighted:
- Reset Highlights

Target sentence you highlighted:
- Reset Highlights

Mismatch between source and target side
What Context do Human Translators Use?
What Context do Human Translators Use?
What Context do Human Translators Use? (Pronoun Anaphora Resolution)
What Context do Human Translators Use? (Pronoun Anaphora Resolution)
Have we got her report? Yes, it’s in the infirmary already.
On dispose de son rapport? Oui, [il / elle] est à l’infirmière.
What Context do Human Translators Use? (Pronoun Anaphora Resolution)

Have we got her report? Yes, it’s in the infirmary already.

On dispose de son rapport? Oui, [il / elle] est à l’infirmière.
Have we got her report? It’s important. Yes, it’s in the infirmary already.

On dispose de son rapport? Il est important. Oui, [il / elle] est à l’infirmière.
What Context do Human Translators Use? (Pronoun Anaphora Resolution)

Have we got her report? It’s important. Yes, it’s in the infirmary already.

On dispose de son rapport? Il est important. Oui, [il / elle] est à l’infirmière.
What Context do Human Translators Use? (Pronoun Anaphora Resolution)

Have we got her report? It’s important. Yes, it’s in the infirmary already.

On dispose de son rapport? Il est important. Oui, [il / elle] est à l’infirmière.
What Context do Human Translators Use? (Pronoun Anaphora Resolution)

Have we got her report? It’s important. Yes, it’s in the infirmary already.

On dispose de son rapport? Il est important. Oui, [il / elle] est à l’infirmière.
What Context do Human Translators Use? (Word Sense Disambiguation)
What Context do Human Translators Use? (Word Sense Disambiguation)
What Context do Human Translators Use? (Word Sense Disambiguation)

Your charm is only exceeded by your frankness.

Ton [charme / portebonheur] n’a d’égal que ta franchise.
What Context do Human Translators Use? (Word Sense Disambiguation)

Your charm is only exceeded by your frankness.

Ton [charme / portebonheur] n’a d’égal que ta franchise.
What Context do Human Translators Use? (Word Sense Disambiguation)

Your charm is only exceeded by your frankness.

Ton [charm / portebonheur] n’a d’égal que ta franchise.
Your charm is only exceeded by your frankness.

**SCAT**: Supporting Context for Ambiguous Translations dataset (14K)
Outline

1. What context is useful during translation?

2. Are models paying attention to this context or not?

3. If not, can we encourage them to do so?
Model

Current source sentence

Encoder

Decoder

Current target sentence
Model

5 previous source sentences

Encoder

Decoder

5 previous target sentences

Current source sentence

Current target sentence
Model

Open Subtitles
Have we got her report?
Yes, it’s in the infirmary already.

On dispose de son rapport?
Oui, il est à l’infirmière.
Quantifying **Human-Model Alignment with SCAT**

**En**
Have we got her report?
Yes, it’s in the infirmary already.

**Fr**
On dispose de son rapport?
Oui, il est à l’infirmière.

**En**
Have we got her report?
Yes, it’s in the infirmary already.

**Fr**
On dispose de son rapport?
Oui, il est à l’infirmière.
Quantifying Human-Model Alignment with SCAT
Quantifying Human-Model Alignment with SCAT
Quantifying Human-Model Alignment with SCAT

Sort

1 2

0.5 0.3 0.1 0.1
Alignment Results

![Alignment Results Graph](image)
Alignment Results
Alignment Results
Alignment Results

![Bar chart showing alignment results for Uniform, Enc Self, Dec Cross, and Dec Self.]
Outline

1. What context is useful during translation?

2. Are models paying attention to this context or not?

3. If not, can we encourage them to do so?
Attention Regularization

$$\mathcal{L}_{\text{NLL}}(\theta) = - \sum_{j=1}^{m} \log p_{\theta}(y_j | x, y_{i<j})$$

OpenSubtitles18
Attention Regularization

\[ \mathcal{L}_{NLL}(\theta) = - \sum_{j=1}^{m} \log p_{\theta}(y_j | x, y_{i<j}) \]

OpenSubtitles18

\[ \mathcal{R}(\theta) = -\lambda \text{KL}(\alpha_{\text{human-norm}} \parallel \alpha_{\text{model}(\theta)}) \]

Context-aware MT Model

SCAT
Evaluation

● BLEU

● COMET
Evaluation

- BLEU
- COMET
- Pronouns F-measure
Oui, il est déjà à l’infirmerie.

Oui, elle est déjà à l’infirmerie.
Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>Baseline</th>
<th>Attention-Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>33.5</td>
<td>41.7</td>
</tr>
<tr>
<td>COMET</td>
<td>36</td>
<td>60</td>
</tr>
<tr>
<td>Pronouns</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrastive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Attention-Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>33.5</td>
<td>33.9</td>
</tr>
<tr>
<td>COMET</td>
<td>41.7</td>
<td>42.3</td>
</tr>
<tr>
<td>Pronouns</td>
<td>36</td>
<td>42</td>
</tr>
<tr>
<td>Contrastive</td>
<td>60</td>
<td>69</td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Attention-Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>33.5</td>
<td>33.9</td>
</tr>
<tr>
<td>COMET</td>
<td>41.7</td>
<td>42.3</td>
</tr>
<tr>
<td>Pronouns</td>
<td>36</td>
<td>42</td>
</tr>
<tr>
<td>Contrastive</td>
<td>60</td>
<td>69</td>
</tr>
</tbody>
</table>

**Improvement:**
- BLEU: +9
- COMET: +8
Results

Baseline vs Attention-Reg
Results

![Bar chart showing results for Baseline and Attention-Reg across Enc Self, Dec Cross, and Dec Self categories.](image-url)
Have we got her report?
Yes, it’s in the infirmary already.

Have we got her report?
Yes, it’s in the infirmary already.

On dispose de son rapport?
Oui, elle est déjà à l’infirmerie.

On dispose de son rapport?
Oui, il est déjà à l’infirmerie.
Have we got her report?
Yes, it’s in the infirmary already.

More experiments & results in paper:
➔ Increased usage of **supporting context**
Results

More experiments & results in paper:

➔ Increased usage of **supporting context**

➔ Regularizing **encoder self-attention** contributes the most
More experiments & results in paper:

→ Increased usage of **supporting context**
→ Regularizing **encoder self-attention** contributes the most
→ Little difference in **WSD performance**
Measuring and Increasing Context Usage in Context-Aware Machine Translation

Patrick Fernandes, Kayo Yin, Graham Neubig, André Martins

(ACL 2021)
Introduction

➔ Many approaches have been proposed for context-aware MT
  ◆ Concatenation, Multi-Encoder, Cache-Based, Hierarchical Attention, ...

➔ However most of these approaches fail to outperform baselines on high-resourced scenario! [1]

Measuring Context Usage

We hypothesise that models under-utilize context even when able to use it.
Measuring Context Usage

➔ We hypothesise that models under-utilize context even when able to use it.

➔ However, quantifying context usage is tricky!
Measuring Context Usage

➔ We hypothesise that models under-utilize context even when able to use it.

➔ However, quantifying context usage is tricky!

➔ Can we do better than what we do now?
Measuring Context Usage

Recently, Bugliarello et al. [2] proposed *cross-mutual information* (XMI)

\[
XMI(X \rightarrow Y) = H_{qLM}(Y) - H_{qMT}(Y \| X)
\]

Measuring Context Usage

Recently, Bugliarello et al. [2] proposed *cross-mutual information* (XMI)

$$\text{XMI}(X \rightarrow Y) = H_{q_{LM}}(Y) - H_{q_{MT}}(Y|X)$$

We propose *conditional cross-mutual information* (CXMI)

$$\text{CXMI}(C \rightarrow Y|X) = H_{q_{MT_A}}(Y|X) - H_{q_{MT_C}}(Y|X, C)$$

Measuring Context Usage

Recently, Bugliarello et al. [2] proposed cross-mutual information (XMI)

\[
XMI(X \rightarrow Y) = H_{q_{LM}}(Y) - H_{q_{MT}}(Y|X)
\]

We propose conditional cross-mutual information (CXMI)

\[
CXMI(C \rightarrow Y|X) = H_{q_{MT_A}}(Y|X) - H_{q_{MT_C}}(Y|X, C)
\]

Measuring Context Usage

Recently, Bugliarello et al. [2] proposed cross-mutual information (XMI)

\[ \text{XMI}(X \rightarrow Y) = H_{q_{LM}}(Y) - H_{q_{MT}}(Y || X) \]

We propose conditional cross-mutual information (CXMI)

\[ \text{CXMI}(C \rightarrow Y || X) = H_{q_{MTA}}(Y || X) - H_{q_{MTC}}(Y || X, C) \]

Measuring Context Usage

\[ H_{qM} T_A(Y|X) \]
Measuring Context Usage

\[ H_{QMT_A}(Y|X) \]

Uncertainty over translations given the source

\[ H_{QMT_C}(Y|X, C) \]

Uncertainty over translations given the source AND context
Measuring Context Usage

\[ H_{qM_{TA}}(Y|X) \]

Uncertainty over translations given the source

\[ H_{qM_{TC}}(Y|X, C) \]

Uncertainty over translations given the source AND context

\[ CXMI(C \rightarrow Y|X) \]
Measuring Context Usage

➔ What considerations do we need to have about \( q_{MT_A} \) and \( q_{MT_C} \)
Measuring Context Usage

➔ What considerations do we need to have about $q_{MT_A}$ and $q_{MT_C}$

➔ We want to make sure the difference between both probabilities is due to the context!
Measuring Context Usage

➔ What considerations do we need to have about $q_{MT_A}$ and $q_{MT_C}$

➔ We want to make sure the difference between both probabilities is due to the context!

➔ We set $q_{MT_A} = q_{MT_C} = q_{MT}$
Measuring Context Usage: Experiments
Measuring Context Usage: Experiments

We consider a document-level translation task

- IWSLT2017, for both EN->DE and EN->FR
Measuring Context Usage: Experiments

→ We consider a document-level translation task
  ◆ IWSLT2017, for both EN->DE and EN->FR

→ We consider also a pretrained setting with a large model
  ◆ Pretrained a sentence-level model on Paracrawl
Measuring Context Usage: Experiments

→ We consider a document-level translation task
  ◆ IWSLT2017, for both EN->DE and EN->FR

→ We consider also a pretrained setting with a large model
  ◆ Pretrained a sentence-level model on Paracrawl

→ We also evaluate on contrastive datasets for pronoun resolution and lexical cohesion
Measuring Context Usage: Experiments

→ We consider a single model trained with *dynamic* context size
Measuring Context Usage: Experiments

→ We consider a single model trained with *dynamic* context size

→ We then measure CXMI for different levels of context on both the source and target side
Measuring Context Usage: Experiments

➔ Diminishing increase in usage of context AND target context used more than source
Measuring Context Usage: Experiments

<table>
<thead>
<tr>
<th>Context Size</th>
<th>$R_{pb}$</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.365</td>
<td>0.315</td>
<td>0.206</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.366</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.367</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.366</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

(1) Pronouns EN-DE (2) Pronouns EN-FR (3) Lexical Cohesion EN-FR

$\Rightarrow$ We also find that CXMI correlates with previous measures of context usage
Increasing Context Usage

Can we find ways to increase context usage?
Increasing Context Usage

Can we find ways to increase context usage?

We propose CoWord dropout:

\[ r_t^{(i)} = \text{Bernoulli}(p) \]

\[ \tilde{x}_t^{(i)} = \begin{cases} <\text{mask}> & \text{if } r_t^{(i)} = 1 \\ x_t^{(i)} & \text{otherwise} \end{cases} \]
Increasing Context Usage

→ Can we find ways to increase context usage?

→ We propose CoWord dropout:

\[
\begin{align*}
    r_t^{(i)} &= \text{Bernoulli}(p) \\
    \bar{x}_t^{(i)} &= \begin{cases} 
        \langle \text{mask} \rangle & \text{if } r_t^{(i)} = 1 \\
        x_t^{(i)} & \text{otherwise}
    \end{cases}
\end{align*}
\]

The Church is merciful.

It always welcomes the misguided lamb.
Increasing Context Usage

➔ Can we find ways to increase context usage?

➔ We propose CoWord dropout:

\[ r_t^{(i)} = \text{Bernoulli}(p) \]

\[ \tilde{x}_t^{(i)} = \begin{cases} <\text{mask}> & \text{if } r_t^{(i)} = 1 \\ x_t^{(i)} & \text{otherwise} \end{cases} \]

*The Church is merciful.*

It always welcomes the misguided lamb.

\[ \downarrow \]

*The Church is merciful.*

always welcomes the lamb.
Increasing Context Usage: Experiments
Increasing Context Usage: Experiments

→ CoWord dropout consistently increases context usage on the target side!
Increasing Context Usage: Experiments

→ We want to assess the impact of CoWord on translation quality
Increasing Context Usage: Experiments

➔ We want to assess the impact of CoWord on translation quality

➔ We train four models with increasing levels of CoWord dropout:
Increasing Context Usage: Experiments

➔ We want to assess the impact of CoWord on translation quality

➔ We train four models with increasing levels of CoWord dropout:
  ◆ A baseline model with no context
Increasing Context Usage: Experiments

→ We want to assess the impact of CoWord on translation quality

→ We train four models with increasing levels of CoWord dropout:
  ◆ A baseline model with no context
  ◆ A 1-to-2 model with one target sentence as context
Increasing Context Usage: Experiments

- We want to assess the impact of CoWord on translation quality

- We train four models with increasing levels of CoWord dropout:
  - A baseline model with no context
  - A 1-to-2 model with one target sentence as context
  - A 2-to-2 model with one source and one target sentences as context
Increasing Context Usage: Experiments

→ We want to assess the impact of CoWord on translation quality

→ We train four models with increasing levels of CoWord dropout:
  ◆ A baseline model with no context
  ◆ A 1-to-2 model with one target sentence as context
  ◆ A 2-to-2 model with one source and one target sentences as context
  ◆ A multi-encoder (1-to-2) model
Increasing Context Usage: Experiments

→ We want to assess the impact of CoWord on translation quality

→ We train four models with increasing levels of CoWord dropout:
  ◆ A baseline model with no context
  ◆ A 1-to-2 model with one target sentence as context
  ◆ A 2-to-2 model with one source and one target sentences as context
  ◆ A multi-encoder (1-to-2) model

→ We evaluate on BLEU and COMET
Increasing Context Usage: Experiments

CoWord dropout seems to improve performance in all cases for non-pretrained models.
CoWord dropout marginally helps for most cases in the pretrained setting.
Increasing Context Usage: Experiments

CoWord dropout also helps with the multi-encoder architecture
Increasing Context Usage: Experiments

→ Improvements are even more noticeable when evaluating on contrastive datasets
For the multi-encoder, CoWord dropout is needed to beat the baseline!
Summary
Summary

New methods to measure context usage in MT

- Agreement with human annotations
- Conditional cross mutual information (CXMI)

New training methods to increase context usage in MT

- Supervised attention
- Contextual word dropout

⇒ Code & Data:
  github.com/neulab/contextual-mt
Future Directions

- Can we quantify, on a more fine-grained level, under which circumstances we require context?
- Better understanding of context usage across languages and phenomena
- Better methods to increase or optimize context usage

➡ Code & Data:

github.com/neulab/contextual-mt