Selective Attention for Context-aware Neural Machine Translation

Sameen Maruf†, André F. T. Martins‡, Gholamreza Haffari†

†Faculty of Information Technology, Monash University, Australia
‡Unbabel & Instituto de Telecomunicações, Lisbon, Portugal

NAACL-HLT, Minneapolis, June, 2019
Overview

1. The Whys?
2. Proposed Approach
3. Experiments and Analyses
4. Summary
Overview

1. The Whys?
2. Proposed Approach
3. Experiments and Analyses
4. Summary
Why document-level machine translation?

Most state-of-the-art NMT models translate sentences independently. Discourse phenomena are ignored, e.g., pronominal anaphora and coherence, which may have long-range dependency. Most of the works in document NMT focus on using a few previous sentences as context ignoring the rest of the document. ([Jean et al., 2017, Wang et al., 2017, Bawden et al., 2018, Voita et al., 2018, Tu et al., 2018, Zhang et al., 2018, Miculicich et al., 2018])

The global document context for MT ([Maruf and Haffari, 2018])
Why document-level machine translation?

- Most state-of-the-art NMT models translate sentences independently
Why document-level machine translation?

- Most state-of-the-art NMT models translate sentences independently
- Discourse phenomena are ignored, e.g., pronominal anaphora and coherence, which may have long-range dependency
Why document-level machine translation?

- Most state-of-the-art NMT models translate sentences independently.
- Discourse phenomena are ignored, e.g., pronominal anaphora and coherence, which may have long-range dependency.
- Most of the works in document NMT focus on using a few previous sentences as context ignoring the rest of the document.

[Jean et al., 2017, Wang et al., 2017, Bawden et al., 2018, Voita et al., 2018, Tu et al., 2018, Zhang et al., 2018, Miculicich et al., 2018]
Why document-level machine translation?

- Most state-of-the-art NMT models translate sentences independently
- Discourse phenomena are ignored, e.g., pronominal anaphora and coherence, which may have long-range dependency
- Most of the works in document NMT focus on using a few previous sentences as context ignoring the rest of the document

[Jean et al., 2017, Wang et al., 2017, Bawden et al., 2018, Voita et al., 2018, Tu et al., 2018, Zhang et al., 2018, Miculicich et al., 2018]

- The global document context for MT [Maruf and Haffari, 2018]
Why selective attention for document MT?
Why selective attention for document MT?

- Soft attention over words in the document context
- Forms a long-tail absorbing significant probability mass
- Incapable of ignoring irrelevant words
- Not scalable to long documents
Why selective attention for document MT?

- Soft attention over words in the document context
Why selective attention for document MT?

- Soft attention over words in the document context
- Forms a long-tail absorbing significant probability mass
Why selective attention for document MT?

- Soft attention over words in the document context
- Forms a long-tail absorbing significant probability mass
- Incapable of ignoring irrelevant words
Why selective attention for document MT?

- Soft attention over words in the document context
- Forms a long-tail absorbing significant probability mass
- Incapable of ignoring irrelevant words
- Not scalable to long documents
This Work

We propose a sparse and hierarchical attention approach for document NMT which:

- identifies the key sentences in the global document context, and
- attends to the key words within those sentences
Overview

1. The Whys?
2. Proposed Approach
3. Experiments and Analyses
4. Summary
Hierarchical Selective Context Attention

For each query word:

- $\alpha_s$: attention weights given to sentences in context
- $\alpha_w$: attention weights given to words in context
- $\alpha_{hier}$: re-scaled attention weights of words in context

$V_w$: from words in context
Hierarchical Selective Context Attention

For each query word:

- $\alpha_s$: attention weights given to sentences in context
Hierarchical Selective Context Attention

For each query word:

- \( \alpha_s \): attention weights given to sentences in context
- \( \alpha_w \): attention weights given to words in context

\[ \alpha_{\text{hier}} \] re-scaled attention weights of words in context
Hierarchical Selective Context Attention

For each query word:

\[ \alpha_s: \text{attention weights given to sentences in context} \]
\[ \alpha_w: \text{attention weights given to words in context} \]
\[ \alpha_{hier}: \text{re-scaled attention weights of words in context} \]
Hierarchical Selective Context Attention

For each query word:

\[ \alpha_s: \text{attention weights given to sentences in context} \]

\[ \alpha_w: \text{attention weights given to words in context} \]

\[ \alpha_{hier}: \text{re-scaled attention weights of words in context} \]

\[ V_w: \text{from words in context} \]
Hierarchical Selective Attention over Source Document
Hierarchical Selective Attention over Source Document

1. Sparse sentence-level key matching: identify relevant sentences

\[ Q_s \]: representation of words in current sentence

\[ K_s \]: representation of sentences in context
Hierarchical Selective Attention over Source Document

1. **Sparse sentence-level key matching**: identify relevant sentences

\[ Q_s: \text{representation of words in current sentence} \]

\[ K_s: \text{representation of sentences in context} \]
Hierarchical Selective Attention over Source Document

1. **Sparse sentence-level key matching**: identify relevant sentences

```
Q_s: representation of words in current sentence
K_s: representation of sentences in context
```
Hierarchical Selective Attention over Source Document

2. Sparse word-level key matching: identify relevant words in relevant sentences

\[ Q_w \]: representation of words in current sentence

\[ K_w \]: representation of words in context
Hierarchical Selective Attention over Source Document

2. **Sparse word-level key matching**: identify relevant words in relevant sentences

- **$Q_w$**: representation of words in current sentence
- **$K_w$**: representation of words in context
Hierarchical Selective Attention over Source Document

2. Sparse word-level key matching: identify relevant words in relevant sentences

\[ Q_w: \text{representation of words in current sentence} \]
\[ K_w: \text{representation of words in context} \]
Hierarchical Selective Attention over Source Document

3. Re-scale attention weights
Hierarchical Selective Attention over Source Document

3. Re-scale attention weights
Hierarchical Selective Attention over Source Document

4. Read the word-level values with the attention weights
Hierarchical Selective Attention over Source Document

4. Read the word-level values with the attention weights
Proposed Approach

Hierarchical Selective Attention over Source Document

4. Read the word-level values with the attention weights

Our sparse hierarchical attention module is able to selectively focus on relevant sentences in the document context and then attends to key words in those sentences.
Flat Attention over Source Document
Flat Attention over Source Document

- *Soft sentence-level attention* over all sentences in the document context
Flat Attention over Source Document

- *Soft sentence-level attention* over all sentences in the document context

\[ K, V: \text{representation of sentences in context} \]
Proposed Approach

Flat Attention over Source Document

- *Soft sentence-level attention* over all sentences in the document context

\[ K, V: \text{representation of sentences in context} \]
Flat Attention over Source Document

- *Soft sentence-level attention* over all sentences in the document context

\[ K, V: \text{representation of sentences in context} \]

- Comparison to [Maruf and Haffari, 2018]:

![Diagram showing Flat Context Attention and comparison to multi-head attention]
Flat Attention over Source Document

- **Soft sentence-level attention** over all sentences in the document context

- $K, V$: representation of sentences in context

- Comparison to [Maruf and Haffari, 2018]:
  - multi-head attention
Flat Attention over Source Document

- **Soft sentence-level attention** over all sentences in the document context

\[ K, V: \text{representation of sentences in context} \]

- Comparison to [Maruf and Haffari, 2018]:
  - multi-head attention
  - dynamic
Flat Attention over Source Document

- **Soft word-level attention** over all words in the document context

\[ K, V: \text{representation of words in context} \]
Document-level Context Layer

- Hierarchical selective or Flat
Document-level Context Layer

- Hierarchical selective or Flat
Document-level Context Layer

- Hierarchical selective or Flat
- Monolingual context (source) integrated in encoder
Proposed Approach

Document-level Context Layer

- Hierarchical selective or Flat
- Monolingual context (source) integrated in encoder
- Bilingual context (source & target) integrated in decoder
Our Models and Settings

Our Models:
- Hierarchical Attention over context
  - sparse at sentence-level, soft at word-level
  - sparse at both sentence and word-level
- Flat Attention over context
  - soft at sentence-level
  - soft at word-level
Our Models and Settings

Our Models:

Hierarchical Attention over context
- sparse at sentence-level, soft at word-level
- sparse at both sentence and word-level

Flat Attention over context
- soft at sentence-level
- soft at word-level
Our Models and Settings

Our Models:

- Hierarchical Attention over context
  - sparse at sentence-level, soft at word-level
  - sparse at both sentence and word-level
Our Models and Settings

Our Models:
- Hierarchical Attention over context
  - sparse at sentence-level, soft at word-level
  - sparse at both sentence and word-level
- Flat Attention over context
  - soft at sentence-level
  - soft at word-level
Our Models and Settings

Our Models:
- Hierarchical Attention over context
  - sparse at sentence-level, soft at word-level
  - sparse at both sentence and word-level
- Flat Attention over context
  - soft at sentence-level
  - soft at word-level

Our Settings:
- Offline document MT
- Online document MT
Overview

1. The Whys?
2. Proposed Approach
3. Experiments and Analyses
4. Summary
Experimental Setup

Training/dev/test corpora statistics for En-De:

<table>
<thead>
<tr>
<th>Domain</th>
<th>#Sentences</th>
<th>Document length</th>
</tr>
</thead>
<tbody>
<tr>
<td>TED</td>
<td>0.21M/9K/2.3K</td>
<td>120.89/96.42/98.74</td>
</tr>
<tr>
<td>News</td>
<td>0.24M/2K/3K</td>
<td>38.93/26.78/19.35</td>
</tr>
<tr>
<td>Europarl</td>
<td>1.67M/3.6K/5.1K</td>
<td>14.14/14.95/14.06</td>
</tr>
</tbody>
</table>

Baselines:
- Context-agnostic baselines (RNNSearch, Transformer)
- Local source context baselines for online document MT:
  - [Zhang et al., 2018] & [Miculicich et al., 2018]

Evaluation Metrics: BLEU, METEOR
Bilingual Context integration in Decoder (Online Setting)

BLEU scores for different datasets:
- TED: 23.28
- News: 22.78
- Europarl: 28.72
Bilingual Context integration in Decoder (Online Setting)

![BLEU scores](chart.png)

- Europarl: Transformer 28.72, [Miculicich et al., 2018] 29.58
Bilingual Context integration in Decoder (Online Setting)

![Bar Chart](chart.png)

- TED: BLEU scores for Transformer, [Miculicich et al., 2018], Attention(sent), and Attention(word) are 23.28, 24.02, 24.39, and 24.29, respectively.
- News: BLEU scores are 22.78, 24.17, 24.38, and 24.75.
- Europarl: BLEU scores are 28.72, 29.56, 29.58, and 29.9.
Bilingual Context integration in Decoder (Online Setting)

![BLEU Scores](image)

**TED**
- Transformer: 23.28
- [Miculicich et al., 2018]: 24.29
- Attention(sent): 24.02
- Attention(word): 24.62
- H-Attention(sp-soft): 24.43

**News**
- Transformer: 22.78
- [Miculicich et al., 2018]: 24.38
- Attention(sent): 24.17
- Attention(word): 24.75
- H-Attention(sp-soft): 24.36
- H-Attention(sp-sp): 24.58

**Europarl**
- Transformer: 28.72
- [Miculicich et al., 2018]: 29.58
- Attention(sent): 29.56
- Attention(word): 29.9
- H-Attention(sp-soft): 29.8
- H-Attention(sp-sp): 29.64
Bilingual Context integration in Decoder (Online Setting)

Experiments and Analyses

 TED

 News

 Europarl

BLEU

Transformer

[Miculicich et al., 2018]

Attention(sent)

Attention(word)

H-Attention(sp-soft)

H-Attention(sp-sp)
Bilingual Context integration in Decoder (Online Setting)

- TED
- News
- Europarl

<table>
<thead>
<tr>
<th>BLEU</th>
<th>TED</th>
<th>News</th>
<th>Europarl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>24.39</td>
<td>24.75</td>
<td>29.58</td>
</tr>
<tr>
<td></td>
<td>24.29</td>
<td>24.17</td>
<td>29.56</td>
</tr>
<tr>
<td></td>
<td>24.02</td>
<td>24.17</td>
<td>29.8</td>
</tr>
<tr>
<td></td>
<td>23.28</td>
<td>23.56</td>
<td>29.64</td>
</tr>
<tr>
<td></td>
<td>24.62</td>
<td>24.58</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24.43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Transformer
- Attention(sent)
- Attention(word)
- H-Attention(sp-soft)
- H-Attention(sp-sp)

[Miculicich et al., 2018]
Bilingual Context integration in Decoder (Online Setting)

Experiments and Analyses

BLEU scores for TED, News, and Europarl datasets with different context integration methods and baseline:

- TED:
  - Transformer: 23.28
  - Attention(sent): 24.29
  - Attention(word): 24.39
  - Attention(sent-sent): 24.43

- News:
  - Transformer: 22.78
  - Attention(sent): 24.17
  - Attention(word): 24.38
  - Attention(sent-sent): 24.58

- Europarl:
  - Transformer: 28.72
  - Attention(sent): 29.58
  - Attention(word): 29.64
  - Attention(sent-sent): 29.8

**Notes:**
- Attention(sent-sent) refers to H-Attention(sp-soft).
- Attention(sent-sent) refers to H-Attention(sp-sp).
Automatic evaluation metrics for translation do not assess how well models translate inter-sentential phenomena.
Automatic evaluation metrics for translation do not assess how well models translate inter-sentential phenomena.

Measure accuracy of translating English pronoun *it* to its German counterparts *es*, *er* and *sie* using a contrastive test set [Müller et al., 2018].
Analyses

- Automatic evaluation metrics for translation do not assess how well models translate inter-sentential phenomena
- Measure accuracy of translating English pronoun *it* to its German counterparts *es*, *er* and *sie* using a contrastive test set [Müller et al., 2018]
- Perform subjective evaluation in terms of adequacy and fluency [Läubli et al., 2018]
Accuracy of pronoun translation vs. antecedent distance
Accuracy of pronoun translation vs. antecedent distance

Experiments and Analyses

Accuracy

Transformer

0.59

0

0.64

>3

antecedent distance
Accuracy of pronoun translation vs. antecedent distance

- Transformer
- [Miculicich et al., 2018]
- Attention(sent)
- Attention(word)
- H-Attention(sp-soft)
- H-Attention(sp-sp)

Accuracy distribution for different models and antecedent distances.
Accuracy of pronoun translation vs. antecedent distance

<table>
<thead>
<tr>
<th>Model</th>
<th>0 Distance</th>
<th>&gt;3 Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer</td>
<td>0.59</td>
<td>0.64</td>
</tr>
<tr>
<td>Miculicich et al., 2018</td>
<td>0.72</td>
<td>0.66</td>
</tr>
<tr>
<td>Attention(sentence)</td>
<td>0.73</td>
<td>0.66</td>
</tr>
<tr>
<td>Attention(word)</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td>H-Attention(sentence-softmax)</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>H-Attention(sentence-sentence)</td>
<td></td>
<td>0.66</td>
</tr>
</tbody>
</table>

The diagram shows the accuracy of pronoun translation for different models and antecedent distances.
Accuracy of pronoun translation vs. antecedent distance

<table>
<thead>
<tr>
<th>antecedent distance</th>
<th>Transformer</th>
<th>[Miculicich et al., 2018]</th>
<th>Attention(sent)</th>
<th>Attention(word)</th>
<th>H-Attention(sp-soft)</th>
<th>H-Attention(sp-sp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.59</td>
<td>0.72</td>
<td>0.69</td>
<td>0.69</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>&gt;3</td>
<td>0.64</td>
<td>0.66</td>
<td>0.66</td>
<td>0.66</td>
<td>0.68</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Accuracy vs. antecedent distance
## Model Complexity

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params</th>
<th>Speed (words/sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>Transformer</td>
<td>50M</td>
<td>5100</td>
</tr>
<tr>
<td>+Attention, sentence</td>
<td>53.7M</td>
<td>3750</td>
</tr>
<tr>
<td></td>
<td>53.7M</td>
<td>3100</td>
</tr>
<tr>
<td>+H-Attention</td>
<td>54.2M</td>
<td>2600</td>
</tr>
</tbody>
</table>

[Miculicich et al., 2018]
## Model Complexity

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params</th>
<th>Speed (words/sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>Transformer</td>
<td>50M</td>
<td>5100</td>
</tr>
<tr>
<td>+Attention, <em>sentence</em></td>
<td>53.7M</td>
<td>3750</td>
</tr>
<tr>
<td></td>
<td>53.7M</td>
<td>3100</td>
</tr>
<tr>
<td>+H-Attention</td>
<td>54.2M</td>
<td>2600</td>
</tr>
</tbody>
</table>

[Miculicich et al., 2018]
## Model Complexity

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params</th>
<th>Speed (words/sec.)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
<td>Decoding</td>
</tr>
<tr>
<td>Transformer</td>
<td>50M</td>
<td>5100</td>
<td>86.33</td>
</tr>
<tr>
<td>+Attention, <em>sentence</em> word</td>
<td>53.7M</td>
<td>3750</td>
<td>83.84</td>
</tr>
<tr>
<td>+H-Attention</td>
<td>54.2M</td>
<td>2600</td>
<td>74.11</td>
</tr>
</tbody>
</table>

[ Miculicich et al., 2018 ]
## Model Complexity

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params</th>
<th>Speed (words/sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
</tr>
<tr>
<td>Transformer</td>
<td>50M</td>
<td>5100</td>
</tr>
<tr>
<td>+Attention, <em>sentence</em></td>
<td>53.7M</td>
<td>3750</td>
</tr>
<tr>
<td>word</td>
<td>53.7M</td>
<td>3100</td>
</tr>
<tr>
<td>+H-Attention</td>
<td>54.2M</td>
<td>2600</td>
</tr>
<tr>
<td>[Miculicich et al., 2018]</td>
<td>54.8M</td>
<td>1650</td>
</tr>
</tbody>
</table>
## Model Complexity

<table>
<thead>
<tr>
<th>Model</th>
<th>#Params</th>
<th>Speed (words/sec.)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
<td>Decoding</td>
<td></td>
</tr>
<tr>
<td>Transformer</td>
<td>50M</td>
<td>5100</td>
<td>86.33</td>
<td></td>
</tr>
<tr>
<td>+Attention, sentence</td>
<td>53.7M</td>
<td>3750</td>
<td>83.84</td>
<td></td>
</tr>
<tr>
<td>word</td>
<td>53.7M</td>
<td>3100</td>
<td>81.38</td>
<td></td>
</tr>
<tr>
<td>+H-Attention</td>
<td>54.2M</td>
<td>2600</td>
<td>74.11</td>
<td></td>
</tr>
<tr>
<td>[Miculicich et al., 2018]</td>
<td>54.8M</td>
<td>1650</td>
<td>76.90</td>
<td></td>
</tr>
</tbody>
</table>
## Qualitative Analysis

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Src:</strong></td>
<td>Croatia is <strong>their</strong> homeland, too.</td>
</tr>
<tr>
<td><strong>Tgt:</strong></td>
<td>Kroatien ist auch <strong>ihrer</strong> Heimat.</td>
</tr>
<tr>
<td><strong>Transformer:</strong></td>
<td>Kroatien ist auch <strong>seine</strong> Heimat.</td>
</tr>
<tr>
<td><strong>Our Model:</strong></td>
<td>Kroatien ist auch <strong>ihre</strong> Heimatland.</td>
</tr>
</tbody>
</table>
Qualitative Analysis

<table>
<thead>
<tr>
<th>Src: Croatia is their homeland, too.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tgt: Kroatien ist auch ihre Heimat</td>
</tr>
<tr>
<td>Transformer: Kroatien ist auch seine Heimat</td>
</tr>
<tr>
<td>Our Model: Kroatien ist auch ihr Heimatland</td>
</tr>
</tbody>
</table>

Head 8: Top sentences with attention to words related to the antecedent

$s^{j-1}$: to name but a few, these include cooperation with the Hague Tribunal, efforts made so far in prosecuting corruption, restructuring the economy and finances and greater commitment and sincerity in eliminating the obstacles to the return of Croatia’s Serbian population.

$s^{j-4}$: by signing a border arbitration agreement with its neighbour Slovenia, the new Croatian Government has not only eliminated an obstacle to the negotiating process, but has also paved the way for the resolution of other issues.
Head 8: Top sentences with attention to words related to the antecedent

\[ s^{j-1} \]: to name but a few, these include cooperation with the Hague Tribunal, efforts made so far in prosecuting corruption, restructuring the economy and finances and greater commitment and sincerity in eliminating the obstacles to the return of Croatia’s Serbian population.

\[ s^{j-4} \]: by signing a border arbitration agreement with its neighbour Slovenia, the new Croatian Government has not only eliminated an obstacle to the negotiating process, but has also paved the way for the resolution of other issues.
Experiments and Analyses

Qualitative Analysis

| Src: Croatia is **their** homeland, too. | Tgt: Kroatien ist auch **ihre** Heimat. |
| Transformer: Kroatien ist auch **seine** Heimat. | Our Model: Kroatien ist auch **ihr** Heimatland. |

**Head 8: Top sentences with attention to words related to the antecedent**

$s^{j-1}$: to name but a few, these include cooperation with the Hague Tribunal, efforts made so far in prosecuting corruption, restructuring the economy and finances and greater commitment and sincerity in eliminating the obstacles to the return of Croatia’s Serbian population.

$s^{j-4}$: by signing a border arbitration agreement with its neighbour Slovenia, the new Croatian Government has not only eliminated an obstacle to the negotiating process, but has also paved the way for the resolution of other issues.
Qualitative Analysis

| Src: Croatia is their homeland, too. |
| Tgt: Kroatien ist auch ihre Heimat. |
| Transformer: Kroatien ist auch seine Heimat. |
| Our Model: Kroatien ist auch ihr Heimatland. |

Head 8: Top sentences with attention to words related to the antecedent $s^{j-1}$: to name but a few, these include cooperation with the Hague Tribunal, efforts made so far in prosecuting corruption, restructuring the economy and finances and greater commitment and sincerity in eliminating the obstacles to the return of Croatia’s Serbian population.

$s^{j-4}$: by signing a border arbitration agreement with its neighbour Slovenia, the new Croatian Government has not only eliminated an obstacle to the negotiating process, but has also paved the way for the resolution of other issues.
Overview

1. The Whys?
2. Proposed Approach
3. Experiments and Analyses
4. Summary
Summary

Proposed a novel and scalable top-down approach to hierarchical attention for document NMT.

Our experiments in two document MT settings show that our approach surpasses context-agnostic and context-aware baselines in majority cases.

Future Work:
Investigate benefits of sparse attention in terms of better interpretability of context-aware NMT models.
Summary

- Proposed a **novel and scalable** top-down approach to **hierarchical attention** for document NMT
- Our experiments in two document MT settings show that our approach surpasses context-agnostic and context-aware baselines in majority cases
Summary

- Proposed a novel and scalable top-down approach to hierarchical attention for document NMT.
- Our experiments in two document MT settings show that our approach surpasses context-agnostic and context-aware baselines in majority cases.

Future Work:
Investigate benefits of sparse attention in terms of better interpretability of context-aware NMT models.


Document Context Neural Machine Translation with Memory Networks.

DyNet: The Dynamic Neural Network Toolkit.

A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation.

# Implementation and Hyperparameters

**Implementation:**
DyNet C++ interface [Neubig et al., 2017], using *Transformer-DyNet* ([https://github.com/duyvuleo/Transformer-DyNet](https://github.com/duyvuleo/Transformer-DyNet))

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Layers</td>
<td>4</td>
</tr>
<tr>
<td>#Heads</td>
<td>8</td>
</tr>
<tr>
<td>Hidden dimensions</td>
<td>512</td>
</tr>
<tr>
<td>Feed-forward layer size</td>
<td>2048</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Adam ($lr=0.0001$)</td>
</tr>
<tr>
<td>Dropout (Base model)</td>
<td>0.1</td>
</tr>
<tr>
<td>Dropout (Document-level model)</td>
<td>0.2</td>
</tr>
<tr>
<td>Label smoothing</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**Src/Tgt vocab sizes:**
TED 17.1k/23.2k, News 16.9k/23.3k, Europarl 16.6k/25.4k (Joint BPE vocab size 30k)
Monolingual Context Integration in Encoder
Bilingual Context Integration in Decoder
Qualitative Analysis

<table>
<thead>
<tr>
<th>Src: my <strong>thoughts</strong> are also with the victims.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref: meine <strong>Gedanken</strong> sind auch bei den Opfern.</td>
</tr>
<tr>
<td>Transformer: ich <strong>denke</strong> auch an die Opfer.</td>
</tr>
<tr>
<td>Our Model: meine <strong>Gedanken</strong> sind auch bei den Opfern.</td>
</tr>
</tbody>
</table>
Qualitative Analysis

<table>
<thead>
<tr>
<th>Src: my <strong>thoughts</strong> are also with the victims.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref: meine <strong>Gedanken</strong> sind auch bei den Opfern.</td>
</tr>
<tr>
<td>Transformer: ich <strong>denke</strong> auch an die Opfer.</td>
</tr>
<tr>
<td>Our Model: meine <strong>Gedanken</strong> sind auch bei den Opfern.</td>
</tr>
</tbody>
</table>

*Head 2: Top sentences with attention to related words*

$s^{j-2}$: (FR) Madam President, many things have already been said, but I would like to echo all the words of sympathy and support that have already been addressed to the peoples of Tunisia and Egypt.

$s^{j+4}$: it must implement a strong strategy towards these countries.

$s^{j-1}$: they are a symbol of hope for all those who defend freedom.