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

Overview

Why document-level machine translation?

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]

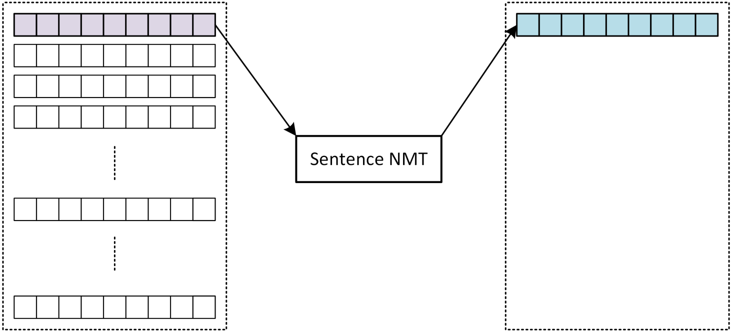
Why document-level machine translation?

- 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

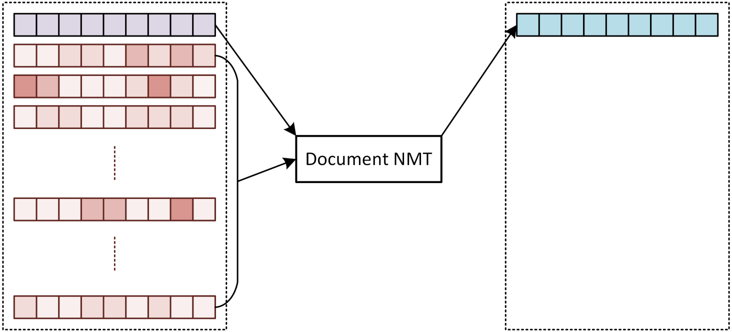
  [Jean et al., 2017, Wang et al., 2017, Bawden et al., 2018, Voita et al., 2018, Tu et al., 2018, Zhang et al., 2018, Miculicich et al., 2018]
- The global document context for MT [Maruf and Haffari, 2018]

Why selective attention for document MT?

Why selective attention for document MT?

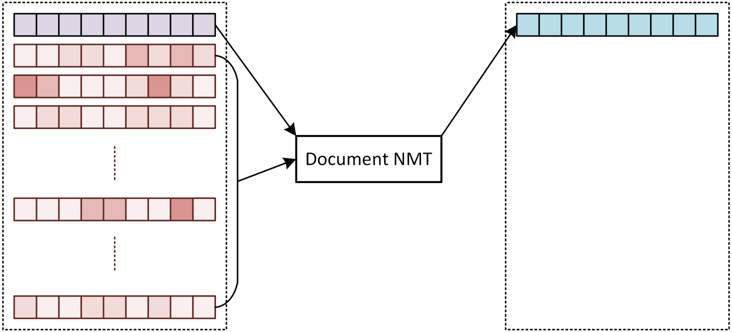


Why selective attention for document MT?



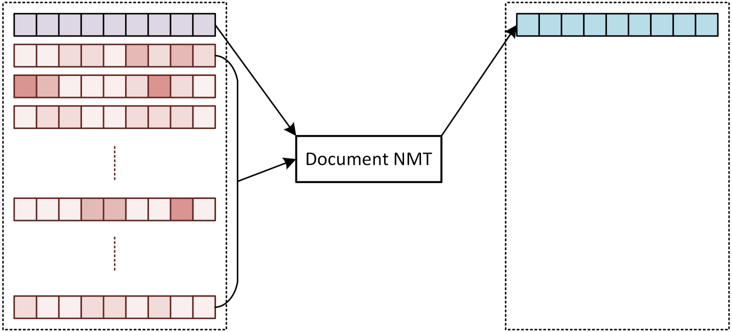
- Soft attention over words in the document context

Why selective attention for document MT?



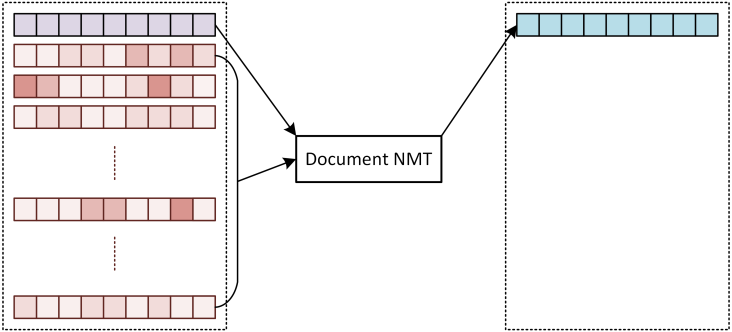
- Soft attention over words in the document context
- Forms a long-tail absorbing significant probability mass

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:

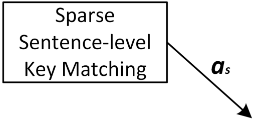
- identifies the key sentences in the global document context, and
- attends to the key words within those sentences

Overview

Hierarchical Selective Context Attention

Hierarchical Selective Context Attention

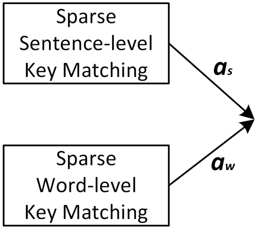
For each query word:



α_s: attention weights given to sentences in context

Hierarchical Selective Context Attention

For each query word:

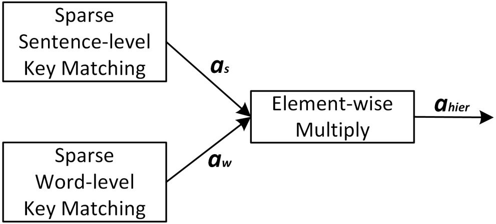


α_s: attention weights given to sentences in context

α_w: attention weights given to words in context

Hierarchical Selective Context Attention

For each query word:



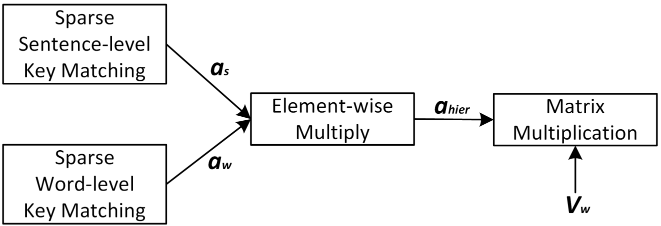
α_s: attention weights given to sentences in context

α_w: attention weights given to words in context

α_{hier}: re-scaled attention weights of words in context

Hierarchical Selective Context Attention

For each query word:



α_s: attention weights given to sentences in context

α_w: attention weights given to words in context

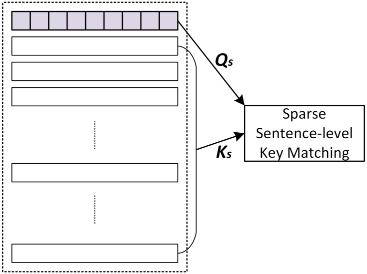
α_{hier}: re-scaled attention weights of words in context

V_w: from words in context

Hierarchical Selective Attention over Source Document

Hierarchical Selective Attention over Source Document

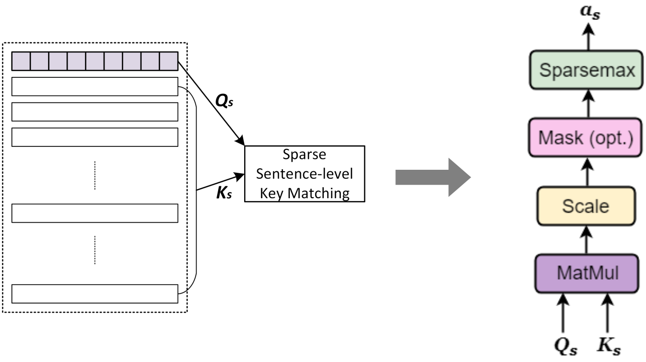
1. *Sparse sentence-level key matching*: identify relevant sentences



$\boldsymbol{Q_s}$: representation of words in current sentence
$\boldsymbol{K_s}$: representation of sentences in context

Hierarchical Selective Attention over Source Document

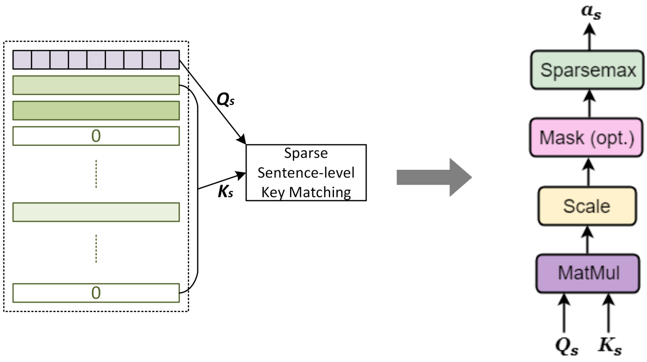
1. *Sparse sentence-level key matching*: identify relevant sentences



Q_s: representation of words in current sentence
K_s: representation of sentences in context

Hierarchical Selective Attention over Source Document

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

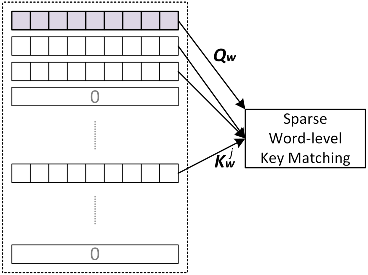


$\boldsymbol{Q_s}$: representation of words in current sentence

$\boldsymbol{K_s}$: representation of sentences in context

Hierarchical Selective Attention over Source Document

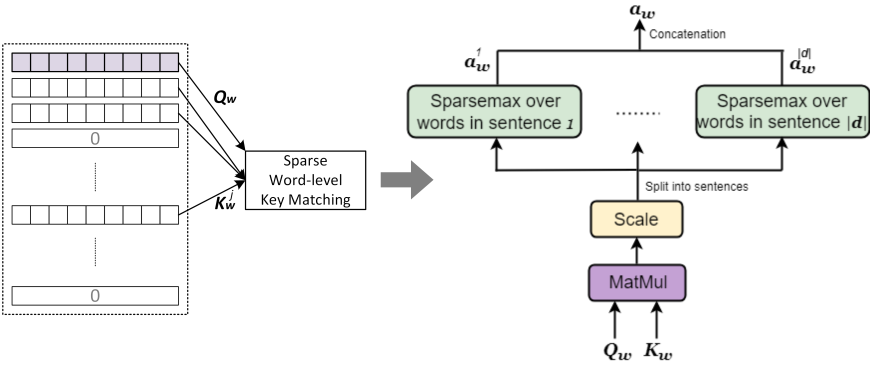
2. *Sparse word-level key matching*: identify relevant words in relevant sentences



Q_w: representation of words in current sentence
K_w: representation of words in context

Hierarchical Selective Attention over Source Document

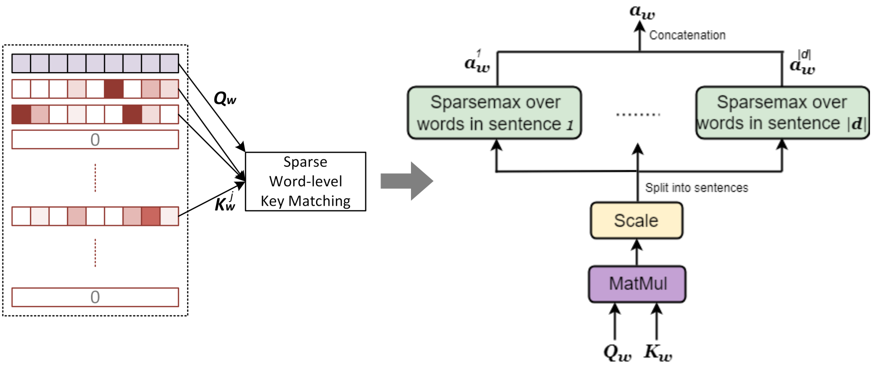
2. *Sparse word-level key matching*: identify relevant words in relevant sentences



Q_w: representation of words in current sentence
K_w: representation of words in context

Hierarchical Selective Attention over Source Document

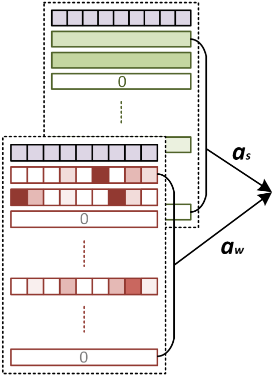
2. *Sparse word-level key matching*: identify relevant words in relevant sentences



Q_w: representation of words in current sentence
K_w: representation of words in context

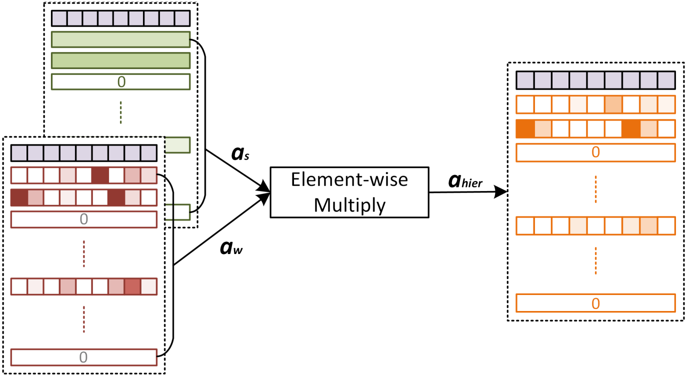
Hierarchical Selective Attention over Source Document

3. *Re-scale attention weights*



Hierarchical Selective Attention over Source Document

3. *Re-scale attention weights*

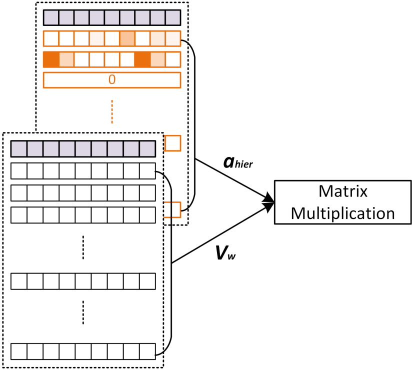


Hierarchical Selective Attention over Source Document

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

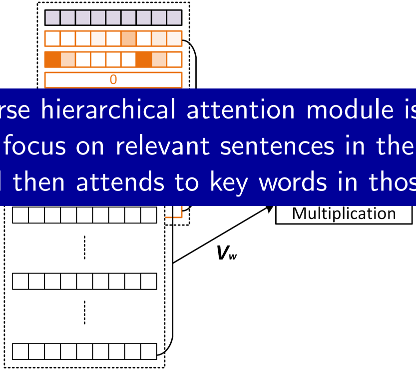
Hierarchical Selective Attention over Source Document

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



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

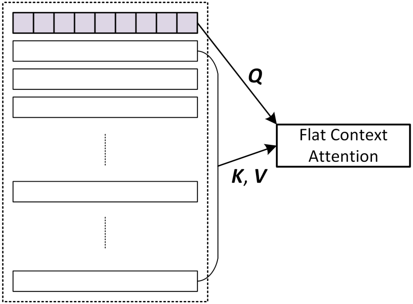
Flat Attention over Source Document

Flat Attention over Source Document

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

Flat Attention over Source Document

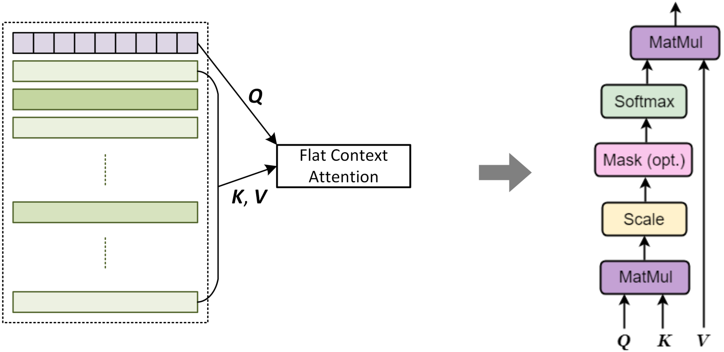
- *Soft sentence-level attention* over all sentences in the document context



$\boldsymbol{K}$, $\boldsymbol{V}$: representation of sentences in context

Flat Attention over Source Document

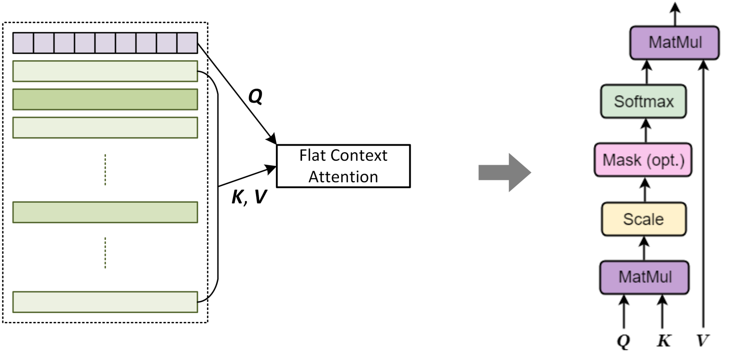
- *Soft sentence-level attention* over all sentences in the document context



$\boldsymbol{K}$, $\boldsymbol{V}$: representation of sentences in context

Flat Attention over Source Document

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

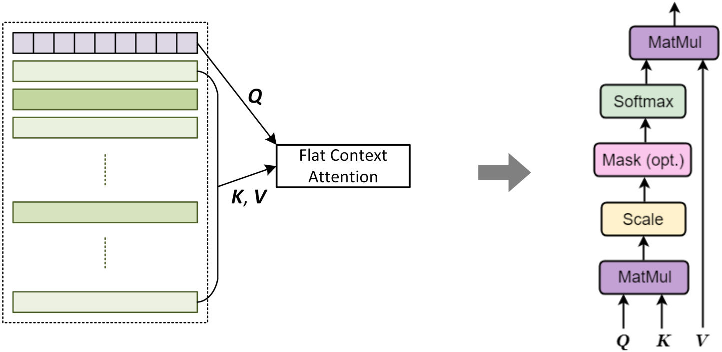


K, ***V***: representation of sentences in context

- Comparison to [Maruf and Haffari, 2018]:

Flat Attention over Source Document

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

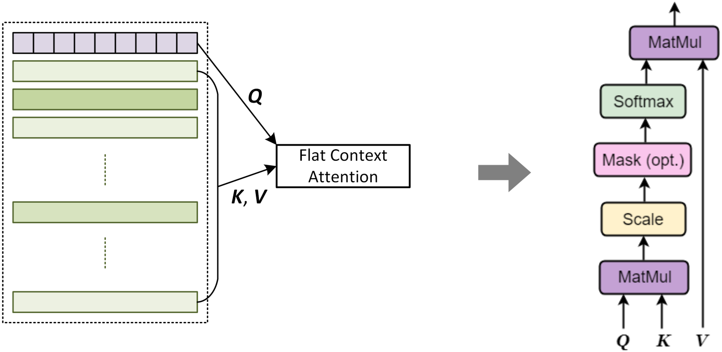


K, ***V***: representation of sentences in context

- 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

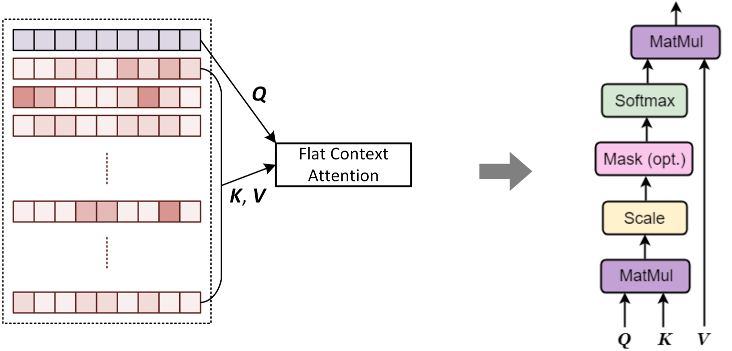


$\boldsymbol{K}$, $\boldsymbol{V}$: representation of sentences in context

- Comparison to [Maruf and Haffari, 2018]:
  - multi-head attention
  - dynamic

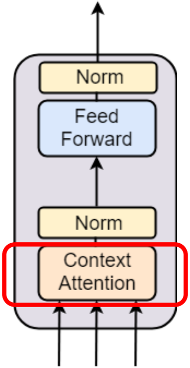
Flat Attention over Source Document

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



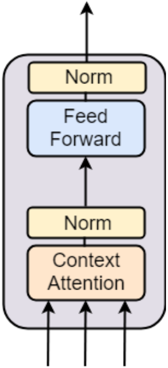
$\boldsymbol{K}$, $\boldsymbol{V}$: representation of words in context

Document-level Context Layer



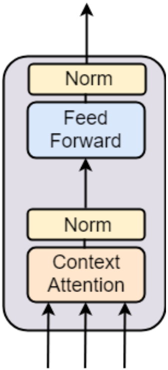
- Hierarchical selective or Flat

Document-level Context Layer



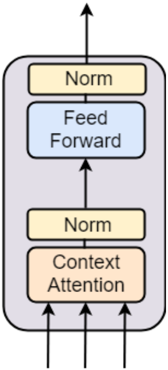
- Hierarchical selective or Flat

Document-level Context Layer



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

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 and Settings

Our Models:

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:

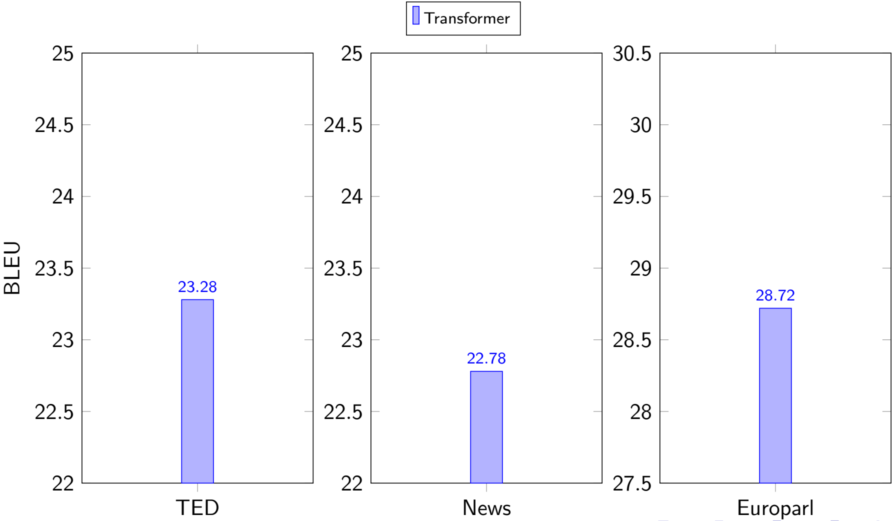
Domain	#Sentences	Document length
TED	0.21M/9K/2.3K	120.89/96.42/98.74
News	0.24M/2K/3K	38.93/26.78/19.35
Europarl	1.67M/3.6K/5.1K	14.14/14.95/14.06

Baselines:

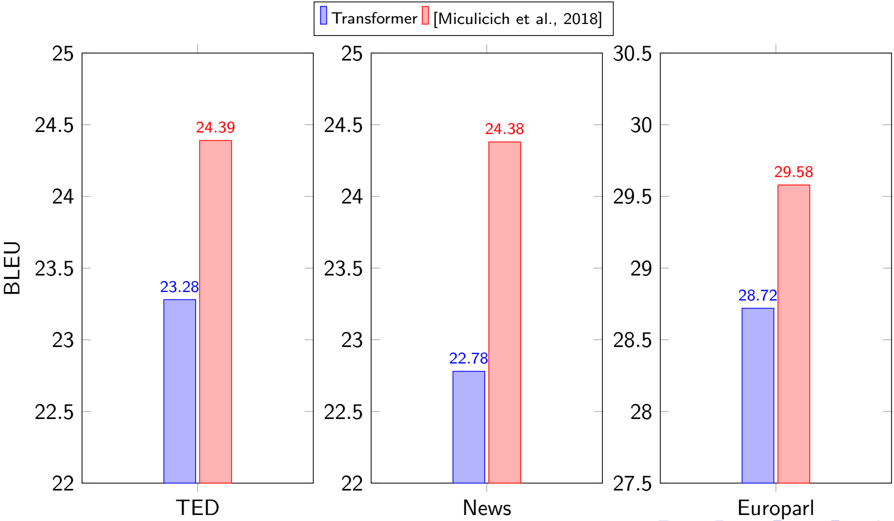
- Context-agnostic baselines (RNNSearch, Transformer)
- Local source context baselines for online document MT:
  - [Zhang et al., 2018] & [Miculicich et al., 2018]

Evaluation Metrics: BLEU, METEOR

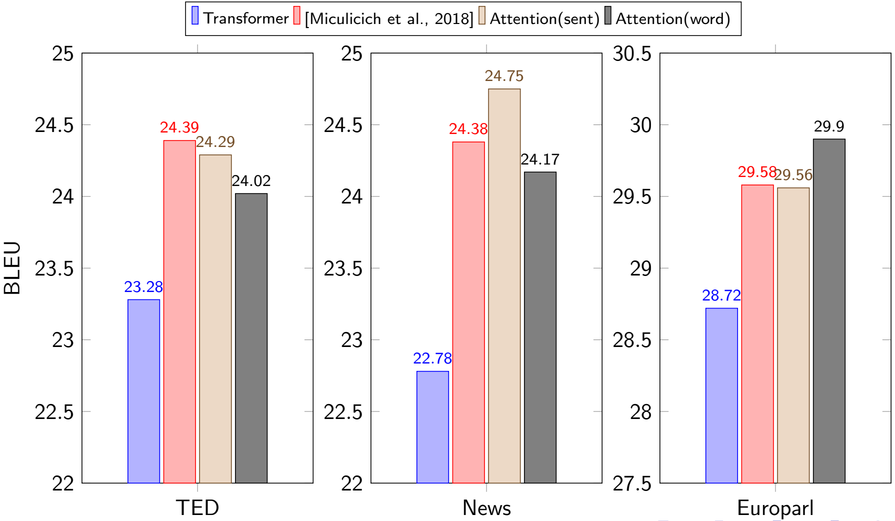
Bilingual Context integration in Decoder (Online Setting)



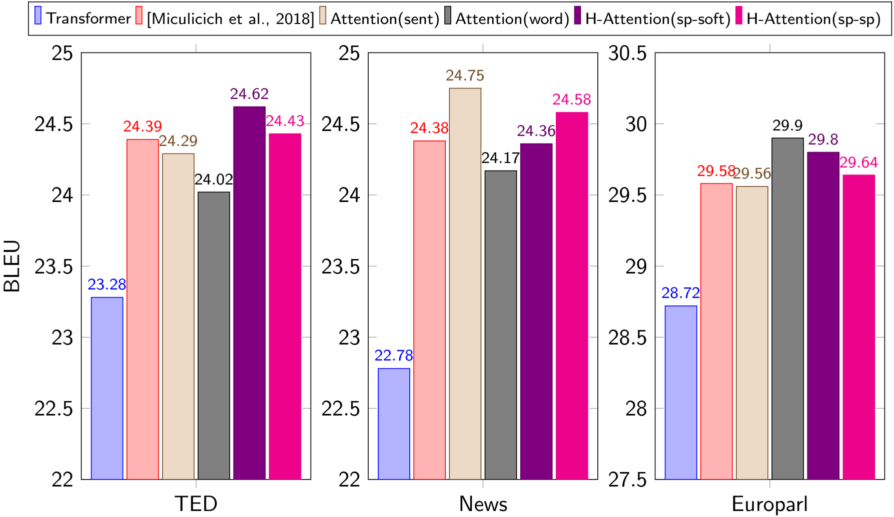
Bilingual Context integration in Decoder (Online Setting)



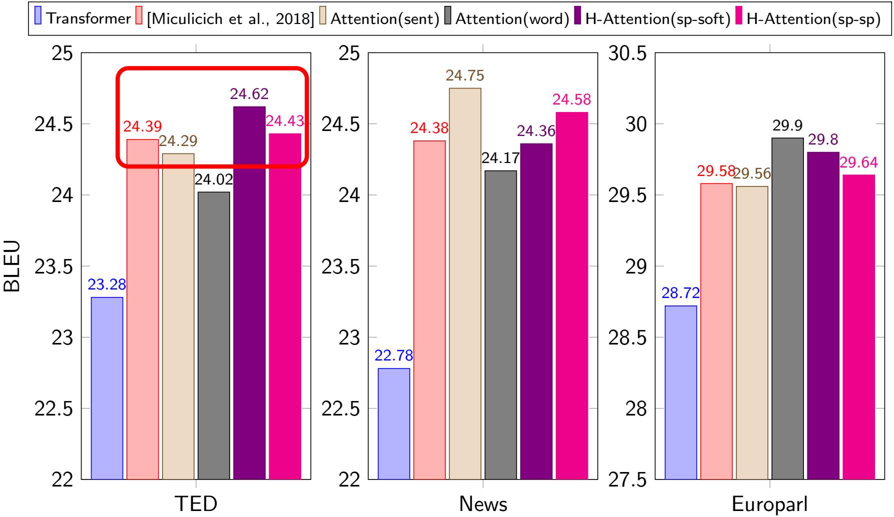
Bilingual Context integration in Decoder (Online Setting)



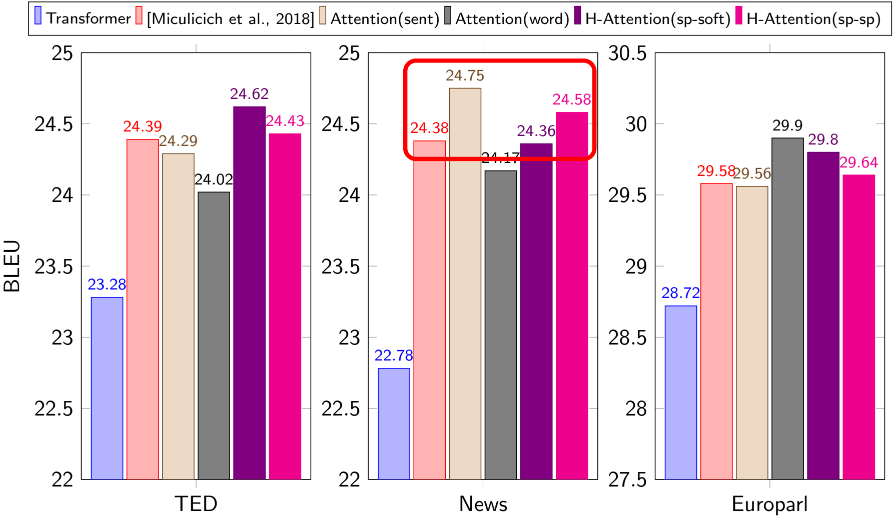
Bilingual Context integration in Decoder (Online Setting)



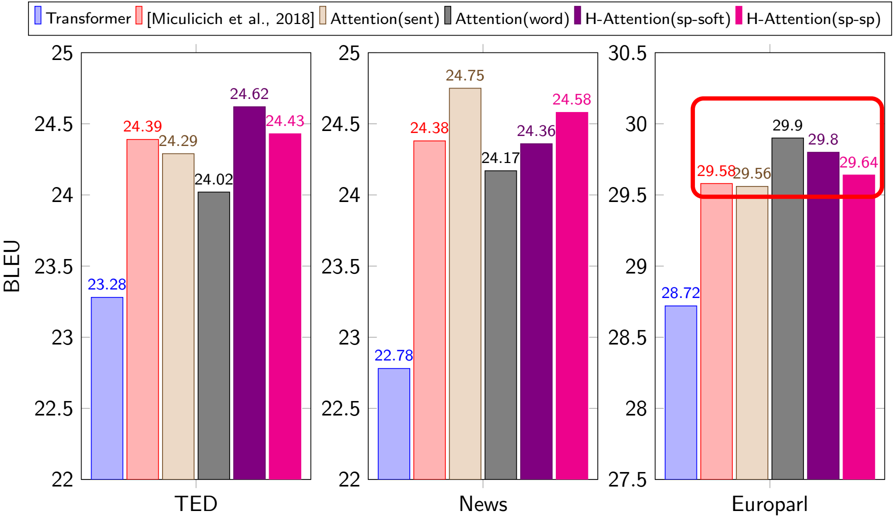
Bilingual Context integration in Decoder (Online Setting)



Bilingual Context integration in Decoder (Online Setting)



Bilingual Context integration in Decoder (Online Setting)



Analyses

- Automatic evaluation metrics for translation do not assess how well models translate inter-sentential phenomena

Analyses

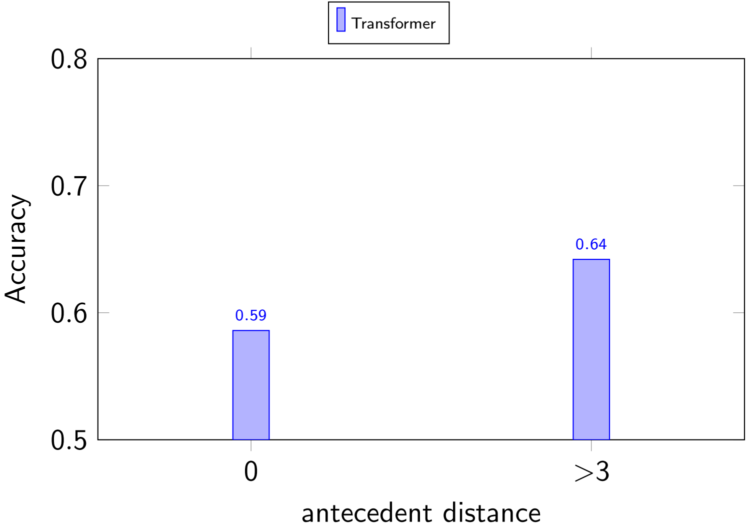
- Automatic evaluation metrics for translation do not assess how well models translate inter-sentential phenomena
- Measure accuracy of translating English pronoun *it* to its German counterparts *es*, *er* and *sie* using a contrastive test set [Müller et al., 2018]

Analyses

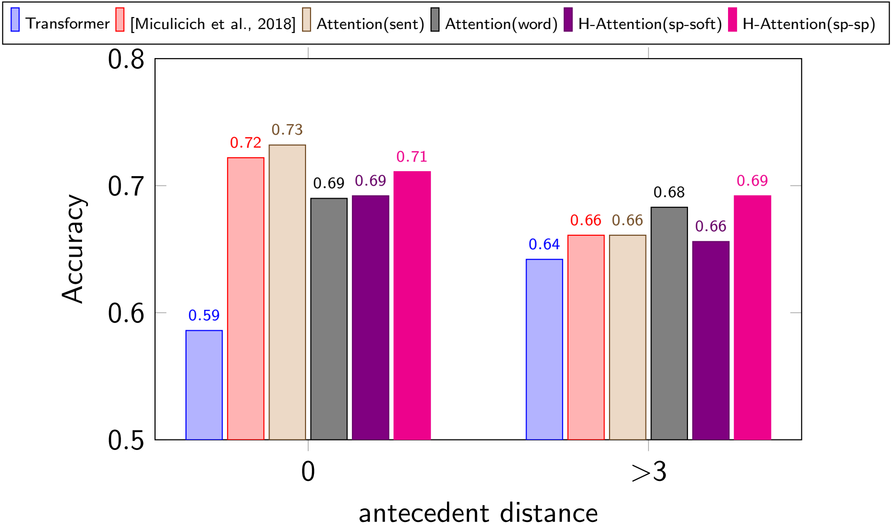
- Automatic evaluation metrics for translation do not assess how well models translate inter-sentential phenomena
- Measure accuracy of translating English pronoun *it* to its German counterparts *es*, *er* and *sie* using a contrastive test set [Müller et al., 2018]
- Perform subjective evaluation in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

Accuracy of pronoun translation vs. antecedent distance

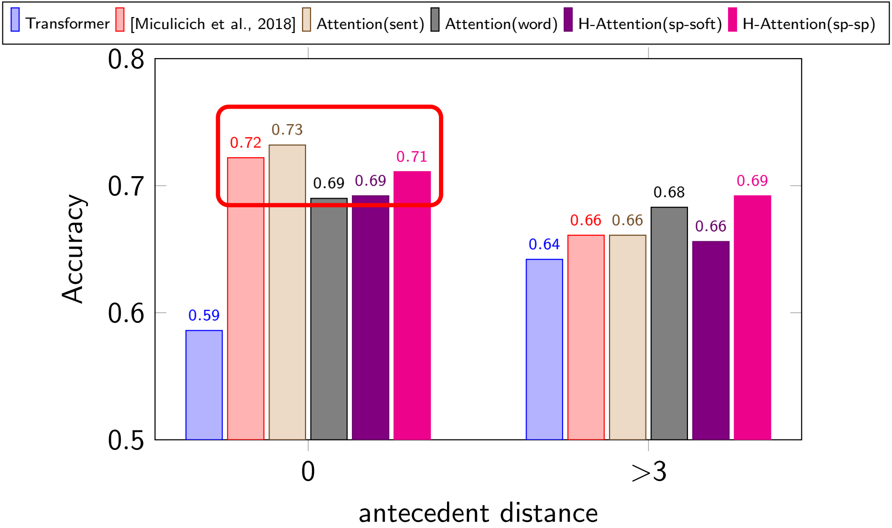
Accuracy of pronoun translation vs. antecedent distance



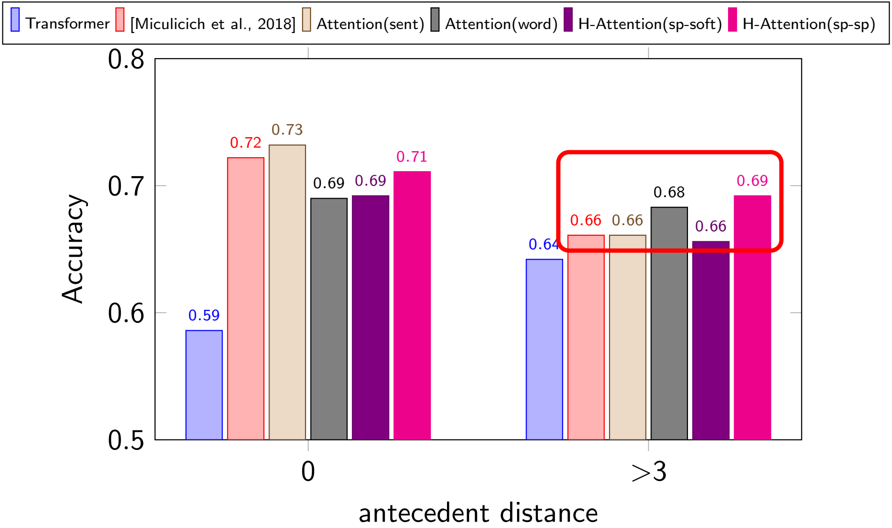
Accuracy of pronoun translation vs. antecedent distance



Accuracy of pronoun translation vs. antecedent distance



Accuracy of pronoun translation vs. antecedent distance



Model Complexity

Model	**#Params**	**Speed (words/sec.)**	
		Training	Decoding
Transformer	50M	5100	86.33
+Attention, *sentence*	53.7M	3750	83.84
word	53.7M	3100	81.38
+H-Attention	54.2M	2600	74.11

Model Complexity

Model	**#Params**	**Speed (words/sec.)**	
		Training	Decoding
Transformer	50M	5100	86.33
+Attention, *sentence*	53.7M	3750	83.84
word	53.7M	3100	81.38
+H-Attention	54.2M	2600	74.11

Model Complexity

Model	**#Params**	**Speed (words/sec.)**	
		Training	Decoding
Transformer	50M	5100	86.33
+Attention, *sentence*	53.7M	3750	83.84
word	53.7M	3100	81.38
+H-Attention	54.2M	2600	74.11

Model Complexity

Model	**#Params**	**Speed (words/sec.)**	
		Training	Decoding
Transformer	50M	5100	86.33
+Attention, *sentence*	53.7M	3750	83.84
word	53.7M	3100	81.38
+H-Attention	54.2M	2600	74.11
[Miculicich et al., 2018]	54.8M	1650	76.90

Model Complexity

Model	**#Params**	**Speed (words/sec.)**	
		Training	Decoding
Transformer	50M	5100	86.33
+Attention, *sentence*	53.7M	3750	83.84
word	53.7M	3100	81.38
+H-Attention	54.2M	2600	74.11
[Miculicich et al., 2018]	54.8M	1650	76.90

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 .	

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

$\boldsymbol{s}^{j-1}$: 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* .

$\boldsymbol{s}^{j-4}$: 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

$\boldsymbol{s}^{j-1}$: 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* .

$\boldsymbol{s}^{j-4}$: 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* .

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

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

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

References I

Jean, S. and Lauly, L. and Firat, O. and Cho, K. (2017).
Does Neural Machine Translation Benefit from Larger Context?
arXiv:1704.05135.

Wang, L. and Tu, Z. and Way, A. and Liu, Q. (2017).
Exploiting Cross-Sentence Context for Neural Machine Translation.
Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Bawden, R. and Sennrich, R. and Birch, A. and Haddow, B. (2018).
Evaluating Discourse Phenomena in Neural Machine Translation.
Proceedings of the NAACL-HLT 2018.

Voita, E. and Serdyukov, P. and Sennrich, R. and Titov, I. (2018).
Context-aware neural machine translation learns anaphora resolution.
Proceedings of ACL 2018.

Tu, Z. and Liu, Y. and Shi, S. and Zhang, T. (2018).
Learning to Remember Translation History with a Continuous Cache.
Proceedings of TACL 2018.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018).
Improving the transformer translation model with document-level context.
Proceedings of EMNLP 2018.

Miculicich, L., Ram, D., Pappas, N., and Henderson, J. (2018).
Document-level neural machine translation with hierarchical attention networks.
Proceedings of EMNLP 2018.

References II

Maruf, S. and Haffari, G. (2018).
Document Context Neural Machine Translation with Memory Networks.
Proceedings of ACL 2018.

Neubig, G. and Dyer, C. and Goldberg, Y. and Matthews, A. and Ammar, W. and Anastasopoulos, A. and Ballesteros, M. and Chiang, D. and Clothiaux, D. and Cohn, T. and Duh, K. and Faruqui, M. and Gan, C. and Garrette, D. and Ji, Y. and Kong, L. and Kuncoro, A. and Kumar, G. and Malaviya, C. and Michel, P. and Oda, Y. and Richardson, M. and Saphra, N. and Swayamdipta, S. and Yin, P. (2017).
DyNet: The Dynamic Neural Network Toolkit.

Müller, M. and Rios, A. and Voita, E. and Sennrich, R. (2018).
A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation.
Proceedings of WMT 2018.

Läubli, S. and Sennrich, R. and Volk, M. (2018).
Has machine translation achieved human parity? A case for document-level evaluation.
Proceedings of EMNLP 2018.

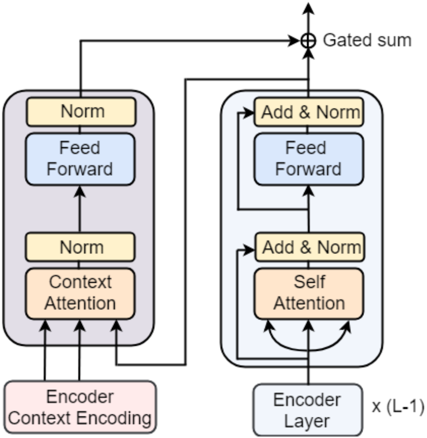
Implementation and Hyperparameters

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

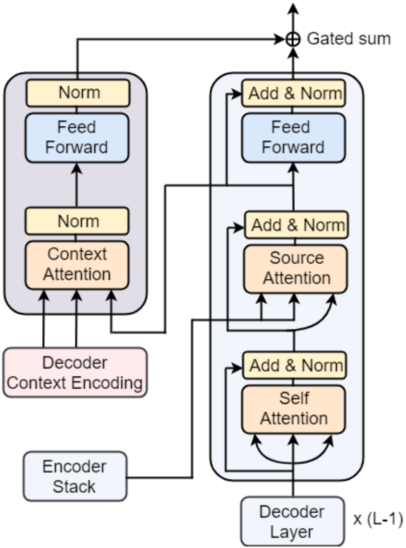
Parameters	Details
#Layers	4
#Heads	8
Hidden dimensions	512
Feed-forward layer size	2048
Optimizer	Adam (*lr*=0.0001)
Dropout (Base model)	0.1
Dropout (Document-level model)	0.2
Label smoothing	0.1

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

Src: my **thoughts** are also with the victims .	
Ref: meine **Gedanken** sind auch bei den Opfern .	
Transformer: ich **denke** auch an die Opfer .	
Our Model: meine **Gedanken** sind auch bei den Opfern .	

Qualitative Analysis

Src: my **thoughts** are also with the victims . Ref: meine **Gedanken** sind auch bei den Opfern .
Transformer: ich **denke** auch an die Opfer . Our Model: meine **Gedanken** sind auch bei den Opfern .

Head 2: Top sentences with attention to related words

$\boldsymbol{s}^{j-2}$: (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 .

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

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