When and Why is Document-level Context Useful in Neural Machine Translation?

Yunsu Kim, Duc Thanh Tran, Hermann Ney

Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany

Lehrstuhl Informatik 6 Human Language Technology and Pattern Recognition

RWTH AACHEN UNIVERSITY

Limits of Document-level Context in NMT

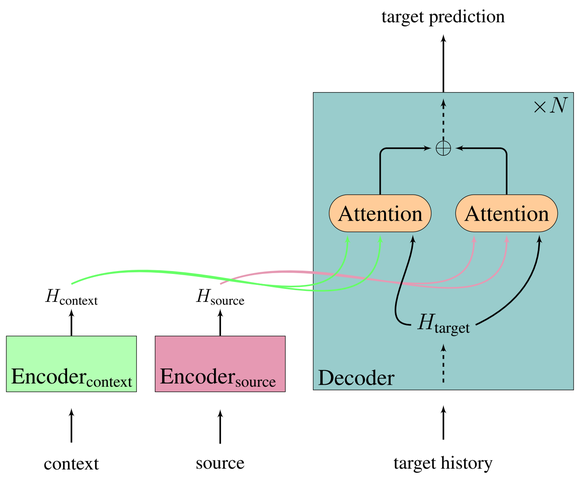
1. How much do we gain from document-level context in NMT?
2. How can we efficiently utilize the context?
3. Are we evaluating it correctly?

Context Encoding Complexity

Question: How much **modeling power** is needed for context encoding?

Analysis: Gradually decrease the depth of context encoder

- Multi-encoder approach with parallel attentions and gating



Encoder$_{\text{context}}$	BLEU [%] en-it	BLEU [%] en-de
None	31.4	28.9
6-layer	**32.7**	30.1
2-layer	32.6	30.2
1-layer	32.2	30.0
Word Embedding	32.5	**30.3**

Answer: Word embeddings are sufficient for context encoding

- No self-attention layers: 22% fewer parameters
- Sequential relation of context tokens is **not** important

Filtering Context Words

Question: Is the **entire context** sentence needed?

Analysis: Filter out specific words in context sentences

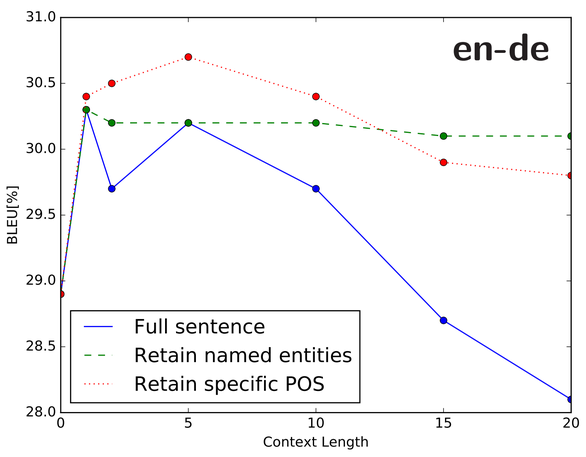
Filtering	Context sentence example	BLEU [%] en-de
None	in recent years, I correctly foresaw that, in the absence of stronger fiscal stimulus (which was not forthcoming in either Europe or the United States)	30.3
Remove stopwords	recent years, I correctly foresaw absence stronger fiscal stimulus (forthcoming Europe United States)	30.3
Remove frequent words	recent correctly foresaw absence stronger fiscal stimulus forthcoming either States	30.2
Retain named entities	recent years Europe the United States	30.3
Retain specific POS	years I foresaw the absence stimulus was forthcoming either Europe or the United States	**30.4**

Answer: Many words in the context sentences are redundant for translation

- Same performance with only 13% of the context words

Context Length

Question: Is a **long-range context** useful?



Analysis: Gradually increase the number of context input sentences

Answer: Only marginal gains

- Training gets unstable
- Word filtering helps
- Worse with >10 sentences

Causes of Improvements

Question: How often is the context utilized for **coreference/coherence**? Is there any **non-linguistic cause** of improvements?

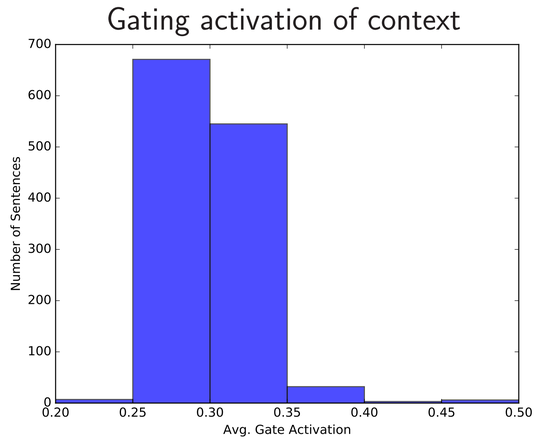
Analysis: Manual inspection of TER-improved hypotheses in the test sets

1. Coreference: gender disambiguation, verb conjugation, ...
2. Topic-aware lexical choice: coherent terminology, style, ...
3. Others: general improvements of fluency/adequacy

	#sentences en-it	#sentences en-de
Total	1,147	2,998
Total TER improved	379	1,246
↪ Coreference	21	2
↪ Topic-aware lexical choice	66	33
↪ Others	292	1,211

Answer: Only **7.5%** of the improvements fix document-level errors

- Others: irrelevant to document-level context
- Does the model actually use the context even in the "irrelevant" cases?



Context is activated in most cases

- No matter how it is utilized
- Constantly allowing additional information flow
  ⇒ Avoid overfitting (**regularization effect**)

Stronger Sentence-level NMT

Question: What if we **minimize the regularization effect** of context?

Analysis: Document-level training on top of a well-regularized system

Training Data	Dropout	System	BLEU [%] en-it	BLEU [%] en-de
Small	0.1	Sentence-level	31.4	28.9
		Document-level	**32.5**	**30.3**
	0.3	Sentence-level	**33.7**	**32.3**
		Document-level	33.5	32.0
Large	0.1	Sentence-level	-	**40.2**
		Document-level	-	39.9

Answer: No improvements in BLEU over strong sentence-level systems

- Targeted test sets: improvements might not carry over to real scenarios

Suggestions

1. **Simplify** the integration of document-level context
2. Do **not** sell the improvements in BLEU by document-level context
3. Check the **real** document-level improvements **manually**
4. Build the sentence-level system **as strong as possible** first

Acknowledgments

This work has received funding from the European Research Council (ERC) (under the European Union's Horizon 2020 research and innovation programme, grant agreement No 694537, project "SEQCLAS"). The work reflects only the authors' views and none of the funding agencies is responsible for any use that may be made of the information it contains.

Links

Paper

Twitter