Document-Level Machine Translation with Large-Scale Public Parallel Corpora

THE UNIVERSITY of EDINBURGH

Proyag Pal **Alexandra Birch** **Kenneth Heafield**

Document-Level MT Needs Document-Level Data

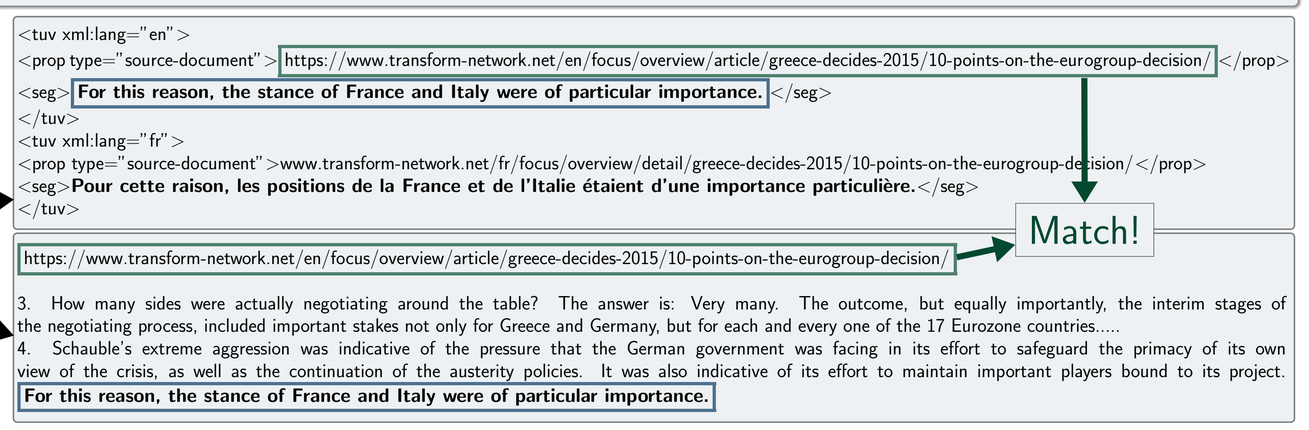
Translation is better when it uses **document context**, but most machine translation (MT) is still **sentence-level**.
This is mostly due to a **lack of document-level parallel data**. Web-crawled parallel corpora usually discard context!
We **recover document context** for ParaCrawl datasets and **release large-scale datasets** to train context-aware MT models.
We evaluate overall MT quality as well as targeted discourse phenomena to confirm that **context-aware models are better**.

Our Dataset

- ParaCrawl aligned webpages, but then only released aligned sentence pairs with source URLs.
- Separately released crawled text with corresponding URLs.
- We **matched the URLs** and recovered up to 512 tokens of **preceding context** for the sentences. A sentence may correspond to multiple URLs, and thus multiple contexts.
- We **release parallel corpora with contexts** for 5 language pairs: eng-deu, eng-fra, eng-ces, eng-pol, eng-rus.

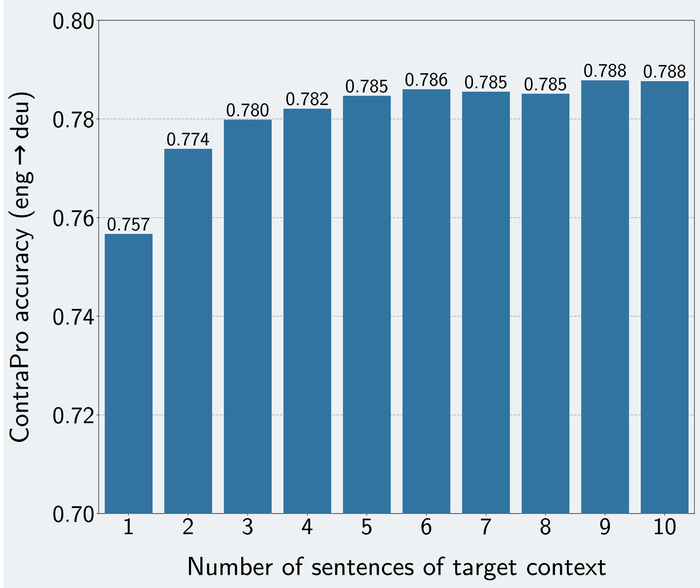
Language pair	Sentence pairs	Source context	Target context	Both contexts
eng-deu	278.3	105.6	110.3	92.1
eng-fra	216.6	83.5	86.3	72.2
eng-ces	50.6	18.7	21.0	16.3
eng-pol	40.1	16.8	18.4	14.9
eng-rus	5.4	3.1	2.8	2.4

Sizes of our document-level datasets in millions of lines.



Contrastive Evaluation

Contrastive evaluations ask models to score alternative translations, only one of which is correct in context.



ContraPro pronoun translation (eng→deu).
Accuracy of sentence-level model: 0.507

Task/Model	→fra	→ces
Lex. Choice:		
Sent.-level	0.5	0.5
Trg. context	**0.525**	**0.533**
Anaphora:		
Sent.-level	0.5	–
Trg. context	**0.545**	–

DiscEvalMT (eng→fra,ces) lexical choice and anaphora resolution.

Model / Context Length	Deixis	Ellipsis		Lex.
		Infl.	VP	Coh.
Sentence-level	0.5	**0.5**	0.058	0.458
Target context – 1 sentence	0.586	**0.5**	0.07	0.468
Target context – 2 sentences	0.654	0.494	0.07	**0.472**
Target context – 3 sentences	**0.692**	**0.5**	**0.074**	**0.472**

GTWiC (eng→rus) with varying amount of target context.

Context-Aware MT Models

We train sentence-level Transformer baselines and dual-encoder Transformers to additionally model context.

Model	BLEU / COMET		
	eng→deu	eng→fra	eng→ces
Sentence-level	35.2 / 85.4	40.5 / 83.1	36.8 / 88.4
Subset - source	35.0 / 85.5	–	36.3 / 87.5
Subset - target	34.3 / 85.3	40.7 / 83.1	35.9 / 87.8
Source context	34.9 / 85.0	–	36.6 / 88.1
Gold target context	**37.4** / **85.9**	**42.6** / **83.2**	**37.3** / **88.5**
Predicted target context	34.7 / 85.4	40.5 / 82.8	35.4 / 87.1

Overall sentence-level BLEU/COMET scores on test sets.
More language pairs in the paper (worse results with smaller datasets).

Contributions & Conclusions

- Released large-scale datasets for document-level MT in several language pairs.
- Context-aware models outperform sentence-level baselines.
- Models improve in terms of overall quality and contrastive evaluation for document-level phenomena.
- Released code to enable the creation of similar corpora for all language pairs supported by ParaCrawl.

More details?

Paper: https://proyag.github.io/files/papers/docmt.pdf
Data: https://hf.co/datasets/Proyag/paracrawl_context
Code: https://github.com/Proyag/ParaCrawl-Context
Contact: proyag.pal@ed.ac.uk

THE UNIVERSITY of EDINBURGH informatics
Edinburgh University of Edinburgh Natural Language Processing NLP
HPLT High Performance Language Technologies
UKRI UK Research and Innovation