Context-aware Decoder for Neural Machine Translation Using a Target-side Document-Level Language Model

Amane Sugiyama

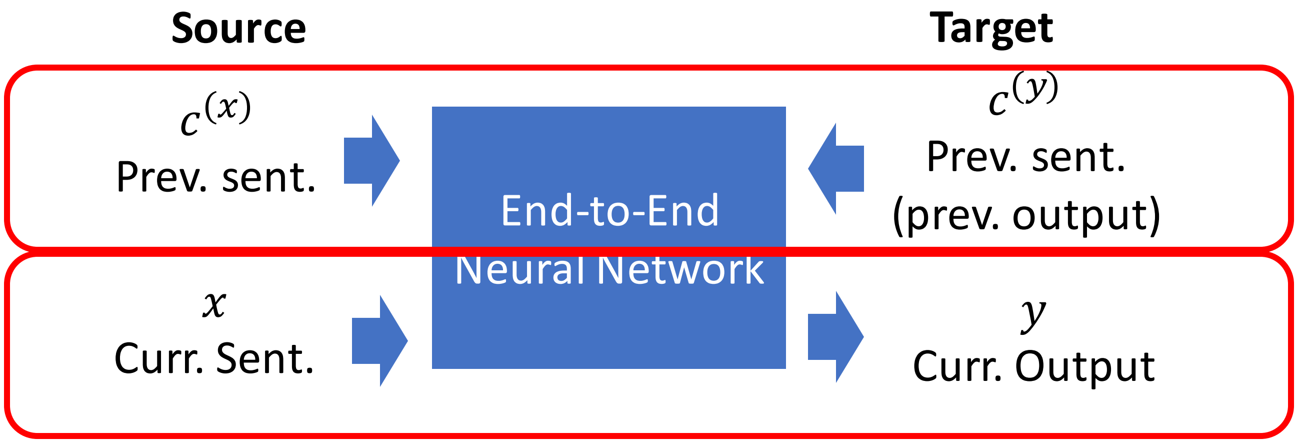
The University of Tokyo

Naoki Yoshinaga

Institute of Industrial Science
The University of Tokyo

Document-level MT and the Standard Approach

Directly **optimize using document-level parallel data**



$$\hat{y} = \arg\max_{y} p\left(y \middle| x, c^{(x)}, c^{(y)}\right)$$

Problem: Lack of document-level parallel data

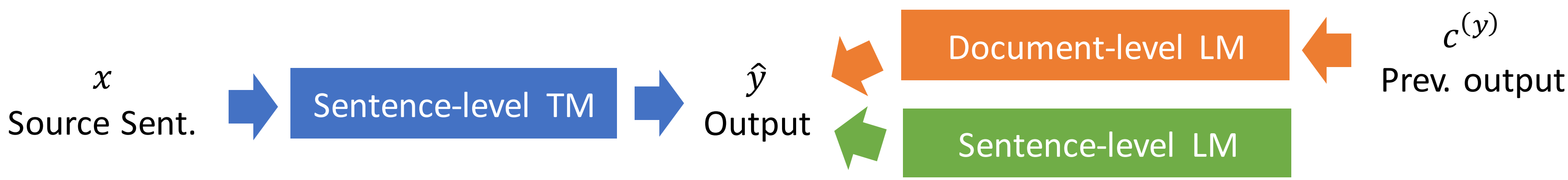
Most of existing parallel data are built from only reliable sentence alignments in parallel/comparable documents.

Can we perform document-level translation without using document-level parallel data?

Decoding with a Document-level Language Model

Approximate the objective function by **sentence-level translation model**, **document-level language model**, and **sentence-level language model** scores.

$$\hat{y} = \arg\max_{y} \log p\left(y \middle| x, c^{(y)}\right) \approx \arg\max_{y} \left[\underbrace{\log p(y|x) + \log p\left(y \middle| c^{(y)}\right) - \log p(y)}_{\textbf{C-Score}} \right]$$



- Document-level parallel is not required for training
- $PMI\left(c^{(y)}, y\right) = \log p\left(y \middle| c^{(y)}\right) - \log p(y)$: association between y and $c^{(y)}$

Decoding Strategy

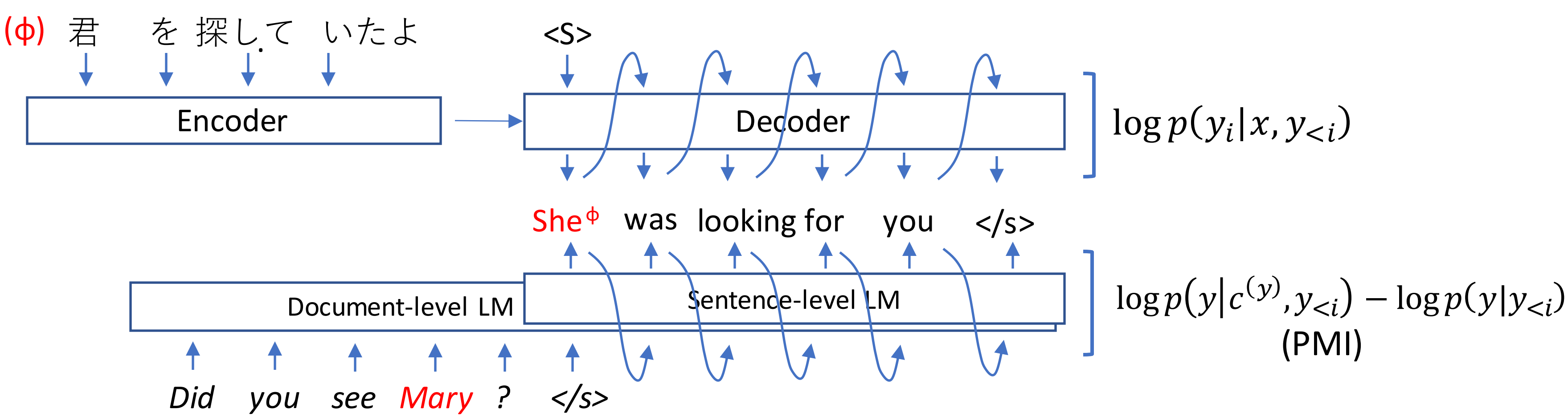
Reranking with C-Score (§2.2.1)

Generate n-best hypotheses by sentence-level decoding and select the one that maximizes C-Score

Context-aware Beam Search (§2.2.2)

Decompose C-Score into token-wise C-Score and perform beam search

$$\mathrm{C{-}Score}(y; x, c^{(y)}) = \sum_i \left[\log p_{S-TM}(y_i | x, y_{<i}) + \log p_{D-LM}(y_i | c^{(y)}, y_{<i}) - \log p_{S-LM}(y_i | y_{<i})\right]$$



Experiments

Overall translation performance measured by BLEU score

Model		para only	+30M mono
Transformer w/ BT	Sentence-level TM	32.36	32.40
DocTransformer	Multi-encoder document-level translation [Zhang+ 18]	**32.50**	31.59
DocRepair	sequence-to-sequence post-editing [Voita+ 18]	n/a	32.35
Bayes DocReranker	Reranking based on scores of S-TM, backward S-TM, and D-LM [Yu+ 20]	n/a	**33.75****
w/o context		n/a	33.67**
Ours (Context-aware beam search)		n/a	32.27
Ours (Reranking with C-Score)		n/a	32.93*

- Bayes DocReranker and ours (rerank) achieved significant improvements the baseline
- Bayes DocReranker performed almost as well without context.

Evaluation of the ability to capture context [Voita+ 2019]

Model	deixis	lex.c	ell.infl	ell.vp
DocTransformer	50.0	45.9	56.0	57.2
DocRepair	**89.1**	75.8	**82.2**	67.2
Bayes DocReranker	65.2	72.2	59.6	44.6
C-Score (ours)	86.9	**94.9**	78.2	**77.0**
PMI	96.8	97.8	75.8	90.6

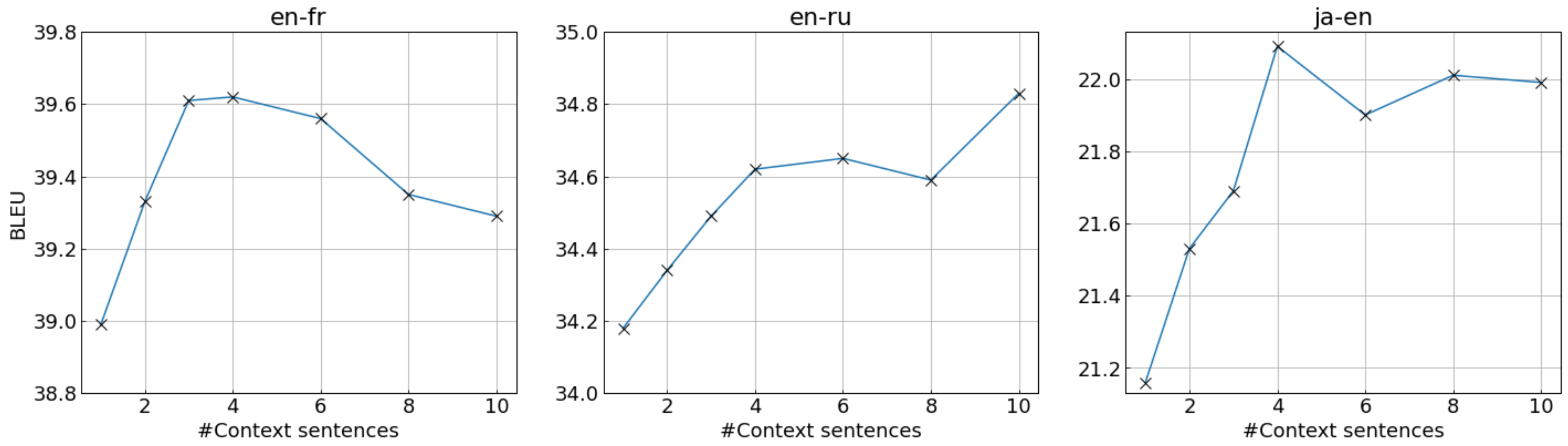
C-Score achieves higher scores than DocRepair in two test sets

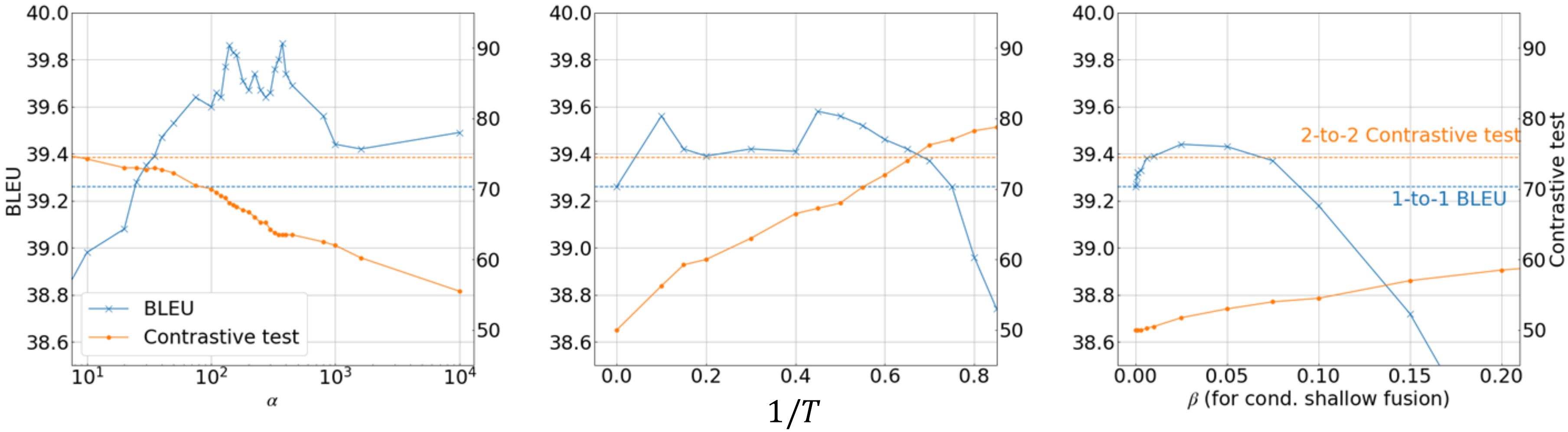
Conclusion

- We proposed an approach to document-level MT, trainable without document-level parallel data
- We confirmed the effectiveness of our methods in terms of BLEU and the contrastive test

Appenix

BLEU vs #context sents.





BLEU
Contrastive test
α
1/T
β (for cond. shallow fusion)
2-to-2 Contrastive test
1-to-1 BLEU