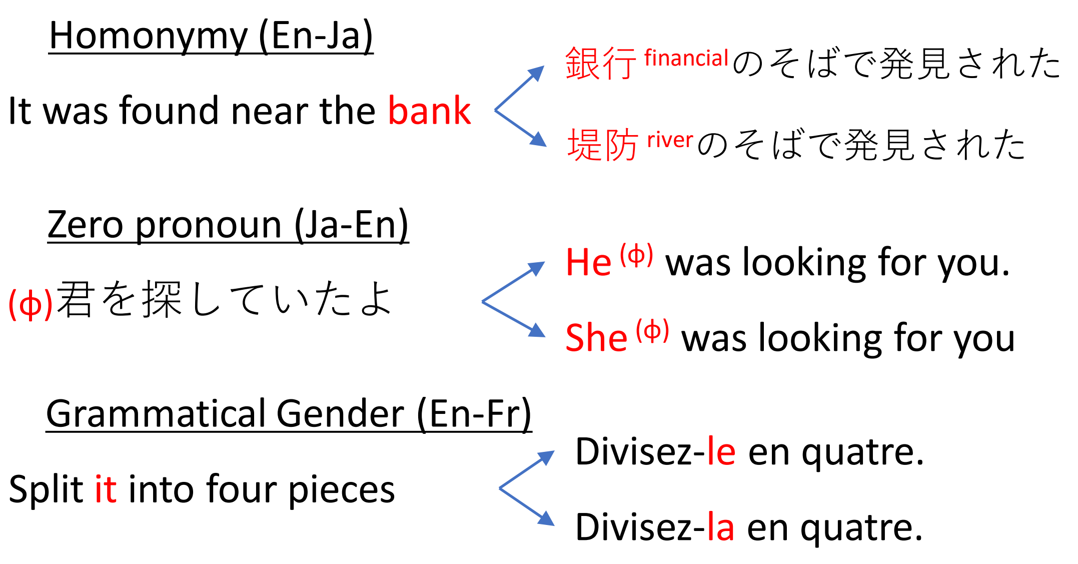
Context-aware Decoder for Neural Machine Translation Using a Target-side Document-Level Language Model

Amane Sugiyama (U.Tokyo, sugi@tkl.iis.u-tokyo.ac.jp) Naoki Yoshinaga (IIS, U.Tokyo, ynaga@iis.u-tokyo.ac.jp)

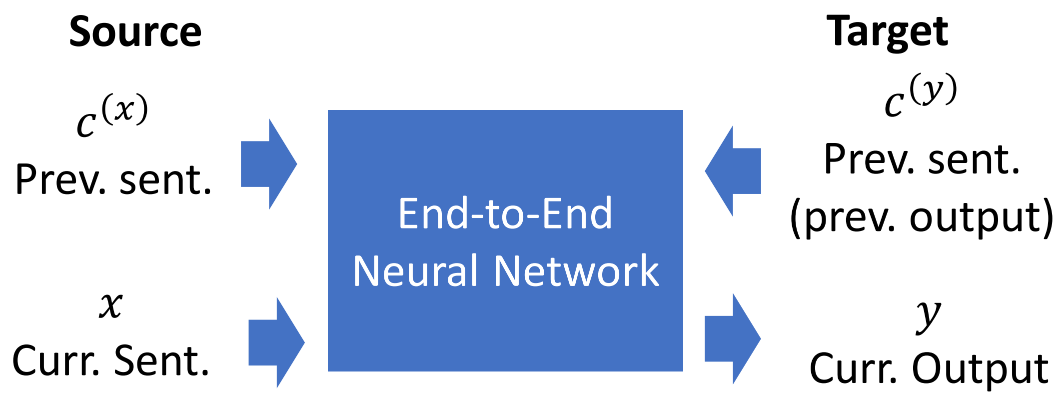
Document-level Machine Translation and the Standard Approach

Context may be necessary for correct translation



Standard approach to context-aware translation

Directly **optimize using parallel data with context** attached (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.

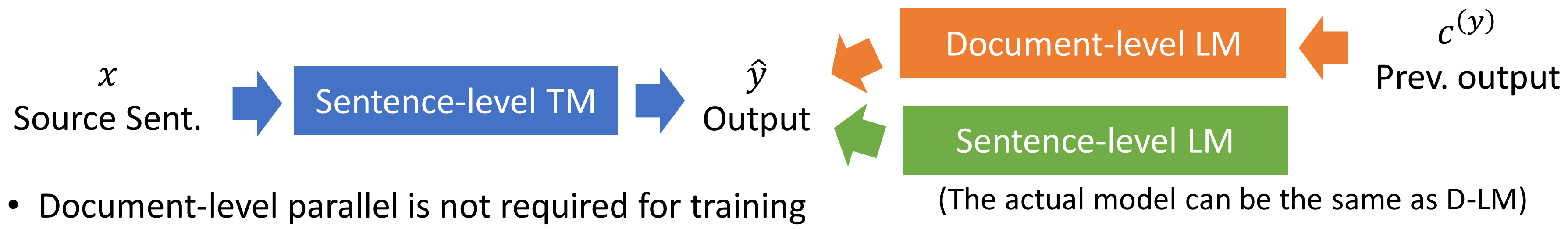
$$\hat{y} = \arg\max_y \log p\left(y \middle| x, c^{(y)}\right) = \arg\max_y \log p\left(c^{(y)} \middle| x, y\right) p(y|x)$$

$$\approx \arg\max_y \log p\left(c^{(y)} \middle| y\right) p(y|x)$$

$$= \arg\max_y \left[\log p(y|x) + \log p\left(y \middle| c^{(y)}\right) - \log p(y)\right]$$

$p\left(c^{(y)} \middle| x, y\right) \approx p\left(c^{(y)} \middle| y\right)$ Assuming x and y are semantically similar

Objective $\text{C–Score}\left(y; x, c^{(y)}\right) = \log p_{S-TM}(y|x) + \log p_{D-LM}\left(y \middle| c^{(y)}\right) - \log p_{S-TM}(y)$

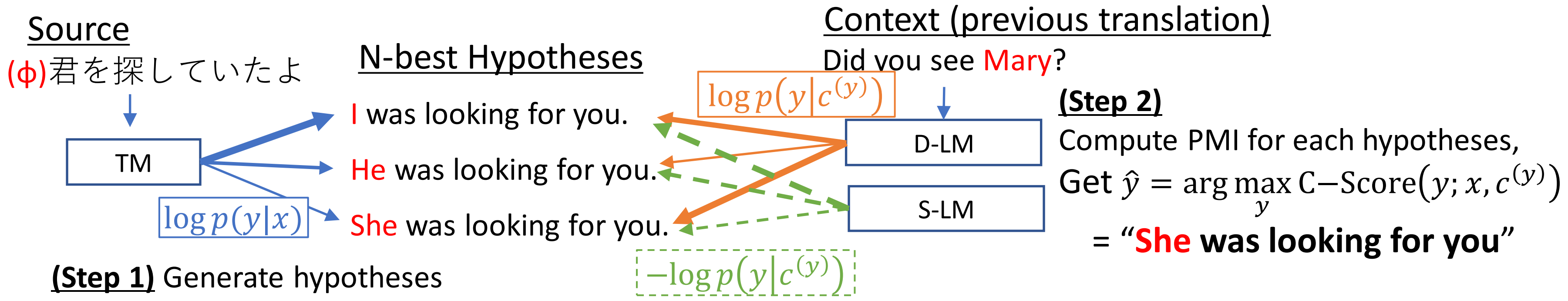


- Document-level parallel is not required for training
- $PMI\left(c^{(y)}, y\right) = \log p\left(y \middle| c^{(y)}\right) - \log p(y)$ represents association between y and $c^{(y)}$
- A hyper param. T [Guo+ 2017] is used to scale PMI: $\text{C–Score} = \log p(y|x) + \left(\log p\left(y \middle| c^{(y)}\right) - \log p(y)\right)/T$

Decoding Strategy

Reranking with C-Score (§ 2.2.1)

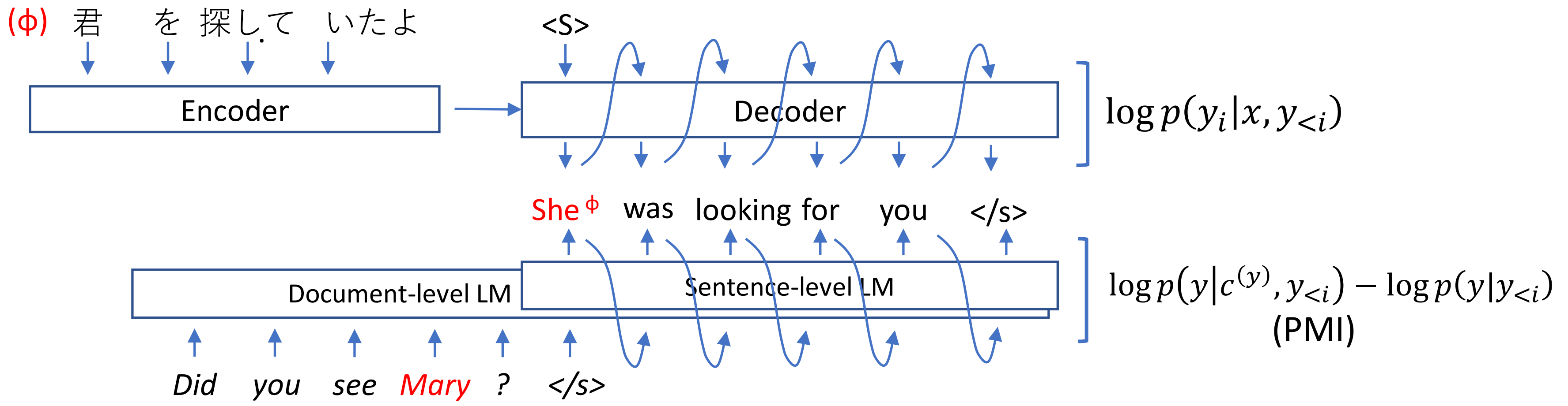
Generate n-best hypotheses by sentence-level decoding and select the one that maximizes C-Score



Context-aware Beam Search (§ 2.2.2)

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

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



The hypotheses for reranking have high reliability but may suffer from the low diversity problem

Experiments

Settings

Data	OpenSubtitles2018 (English -> Russian, parallel: 6M, monolingual: 30M)
Models	Sentence-level TM: Transformer base [Vaswani+ 2017] Document/sentence-level LM: Decoder of Transformer

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*

Only ours (rerank) and Bayes DocReranker achieved significant improvements over Transformer. Bayes DocReranker performed almost as well without context. A brief discussion of inference speed is provided in the paper.

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

Models guess the correct translation out of several candidates based on the translation score

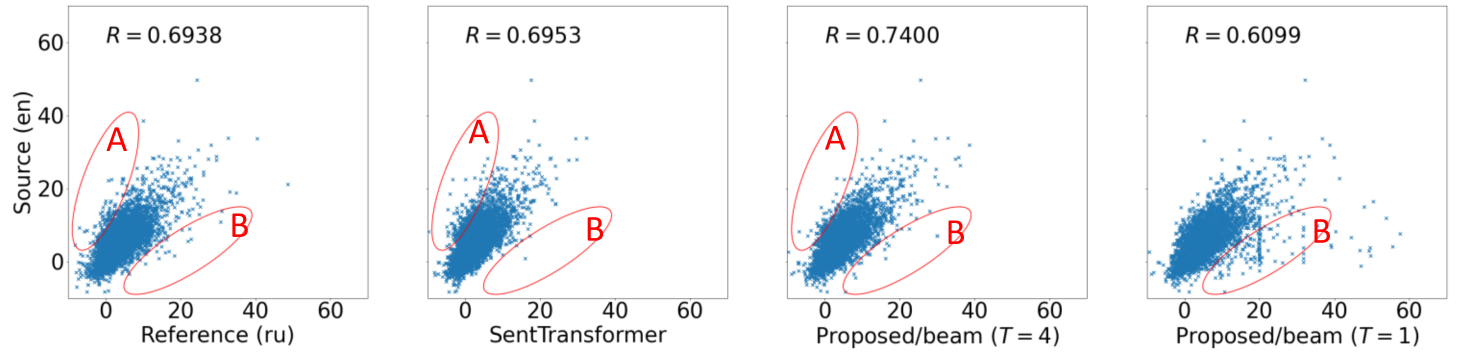
Deixis (person deixis), lex.c (lexical cohesion), ell.infl (inflection of Russian nouns caused by ellipsis), and ell.vp (verb ellipsis in English text not allowed in Russian)

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	86.9	**94.9**	78.2	**77.0**
PMI	96.8	97.8	75.8	90.6

C-Score achieves better lex.c and ell.vp socres and comparable deixis and ell.infl scores to the DocRepair

Analysis: How models change source-target PMI correlation?

Each point stands for a pair of PMI: $\left(\text{PMI}\left(c^{(y_i)}, y_i^{model}\right), \text{PMI}\left(c^{(x_i)}, x_i\right)\right)$ for the i-th sentence pair in a dev set where *model* is the reference y_i and outputs of S-TM, proposed beam search with $T = 0$, and proposed beam search with $T = 4$.



A: Sentence-level Transformer fails to reflect context in translation (low target-side PMI).

B: Decoding with C-Score with T=1 suffers from over-correction

Conclusion

You can perform document-level translation using sentence-level translation model and language model.

Experiment results for En-Fr and En-Ja are available in the v1 paper on arXiv.org (https://arxiv.org/abs/2010.12827v1)