Improving Context-aware Neural Machine Translation with Target-side Context

Hayahide Yamagishi and Mamoru Komachi
Tokyo Metropolitan University, Tokyo, Japan
@PACLING2019, Hanoi, Vietnam

Abstract

- Neural Machine Translation has become popular.
  - It is said NMT can consider sentence-level contexts.

- Recently, some researches expand the context window from sentence-level to document-level

  → Context-aware neural machine translation (CNMT)

- Previous researches of CNMT found the source-side context improves the performance.
  - There are few researches using the target-side context.

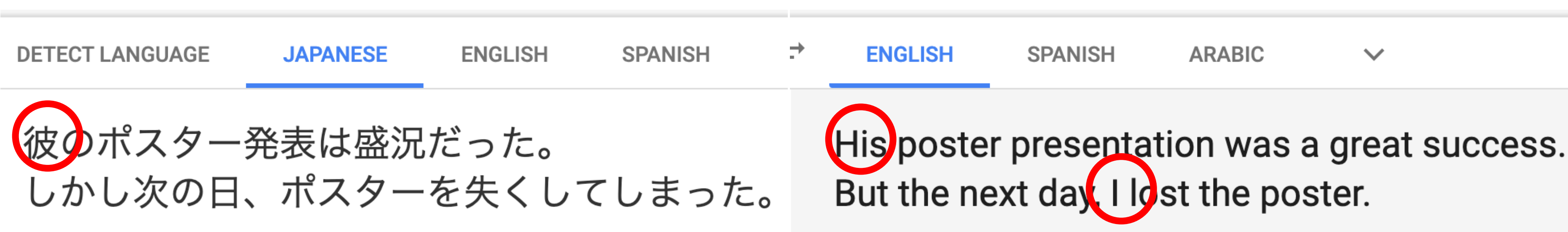
- This research: how can we use the target-side context?

NMT does not employ the context.

- Human translation can use the document context.
  - Can keep the coherence of words and styles.

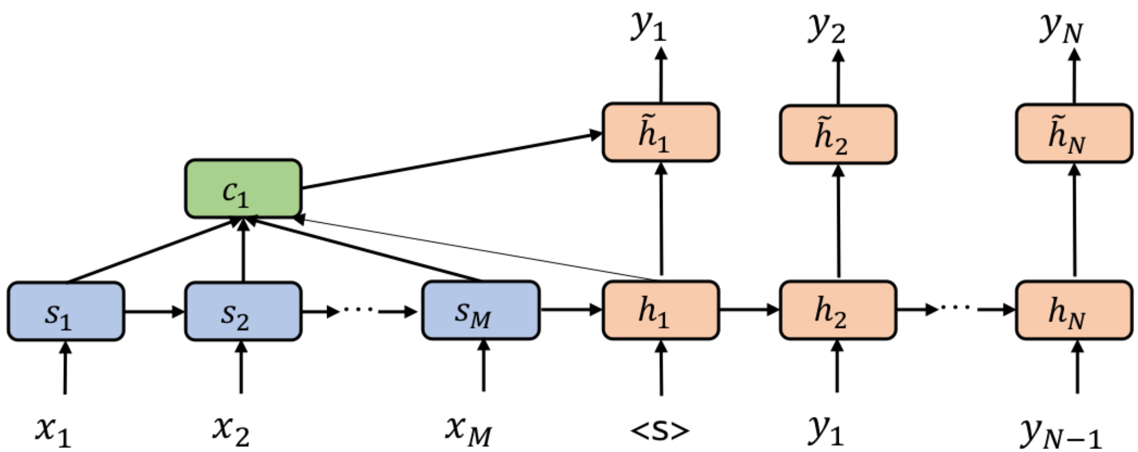
- The sentence in a pro-drop language such as Japanese often omits pronouns if it is apparent from a context.

- Google Translate (used in 9/30/2019)
  - Second sentence mistakes the possessor of "poster", because Japanese sentence omits the subject noun, "彼 (He)".
  - Today's NMT cannot use the inter-sentential context.



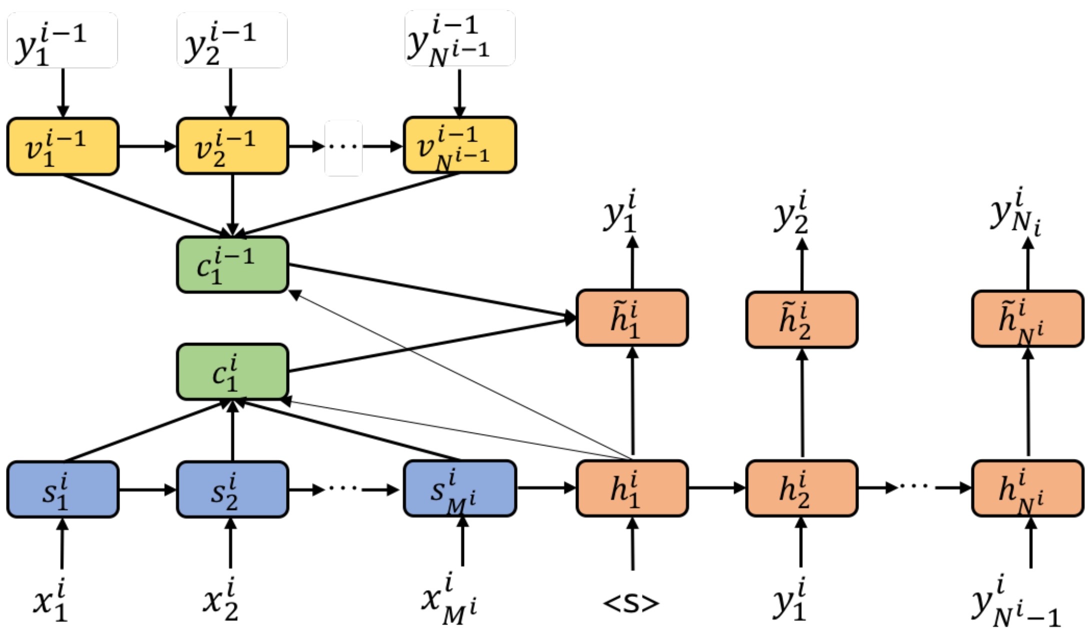
LSTM-based attentional NMT [Luong+, EMNLP15]

- Encoder
  - Change the word sequence x_m into hidden states s_m and sentence representation.
- Attention
  - Calculate the attention vector c_t which refers the source-side information, using the hidden states of the encoder.
- Decoder
  - Generate the word y_t using the attention vector c_t , hidden state of decoder h_t and previous result y_{t-1} .



Multi-Encoder [Bawden+, NAACL18]

- Most popular model of Context-aware NMT [Müller+,WMT18]
  - English-German [Jean+, arXiv 2017.04], [Müller+, WMT18]
  - English-French [Bawden+, NAACL18]
  - English-Russian [Voita+, ACL18]
  - Chinese-English [Zhang+, EMNLP18]
- In this slide, we call this model “Separated model”



Separated model

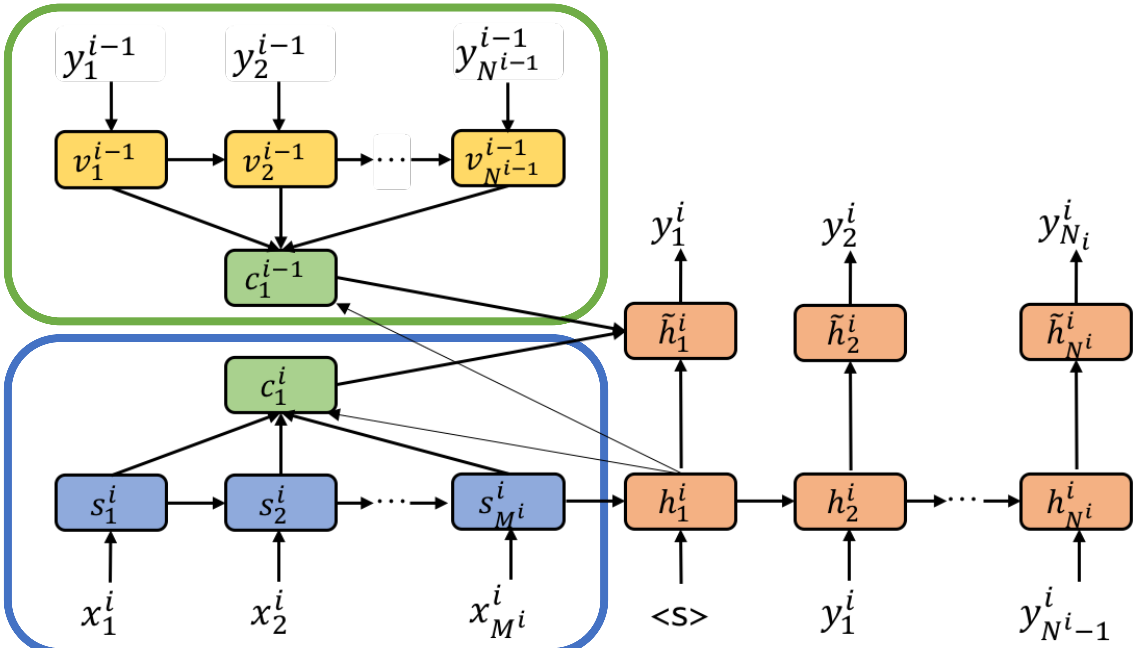
Get the information of two sentence with two encoders.

- Context Encoder: read a previous sentence as a context.

$$v_t^{i-1} = \mathrm{LSTM}_{\mathrm{context}}\left(W_y y_t^{i-1}, v_{t-1}^{i-1}\right)$$

- Encoder: read an input sentence

$$s_m^i = \mathrm{LSTM}_{\mathrm{encoder}}\left(W_x x_m^i, s_{m-1}^i\right)$$



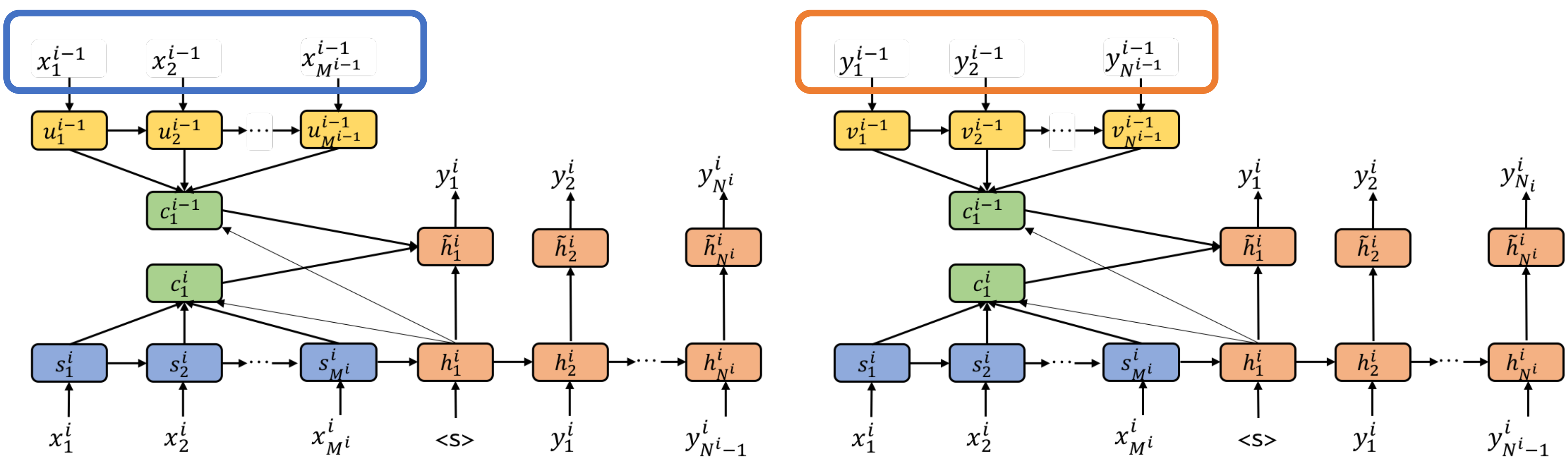
Two patterns of context encoder

- Separated source: use the source-side context.

$$u_t^{i-1} = \mathrm{LSTM}_{\mathrm{context}}\left(\boxed{W_x x_t^{i-1}}, u_{t-1}^{i-1}\right)$$

- Separated target: use the target-side context.

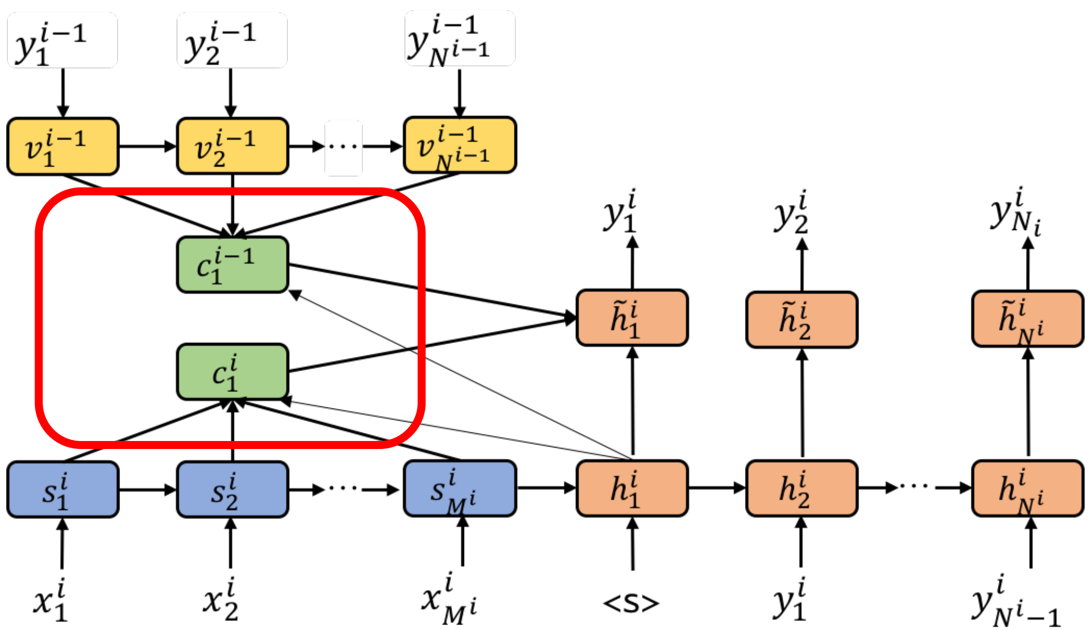
$$v_t^{i-1} = \mathrm{LSTM}_{\mathrm{context}}\left(\boxed{W_y y_t^{i-1}}, v_{t-1}^{i-1}\right)$$



Separated model

The attention $\boldsymbol{c}_n^i$ and context attention $\boldsymbol{c}_n^{i-1}$ vectors are calculated with results of each encoder.

$$\boldsymbol{c}_n^i = \sum_{m=1}^{M^i} \mathrm{softmax}\left(\boldsymbol{s}_m^i \cdot \boldsymbol{h}_n^i\right) \boldsymbol{s}_m^i , \qquad \boldsymbol{c}_n^{i-1} = \sum_{t=1}^{N^{i-1}} \mathrm{softmax}\left(\boldsymbol{v}_t^{i-1} \cdot \boldsymbol{h}_n^i\right) \boldsymbol{v}_t^{i-1}$$

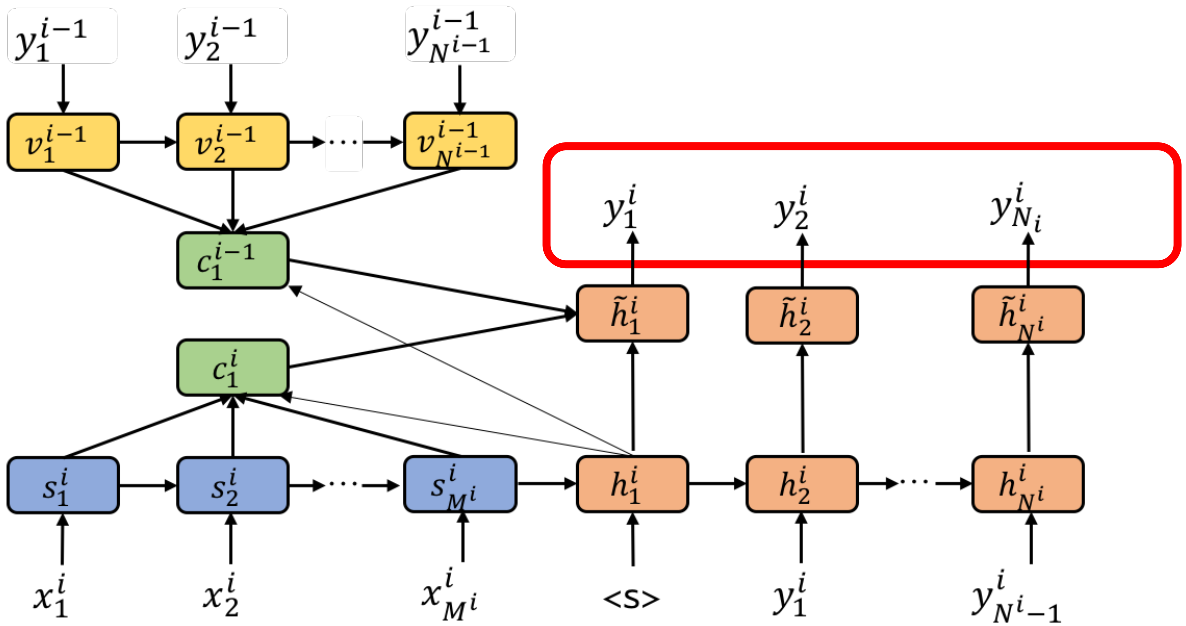


Separated model

- The objective is maximizing p for i = 1, ⋯ L, .

$$p\left(Y^i \middle| X^i, Z^{i-1}\right) = \prod_{n=1}^{N^i} p\left(y_n^i \middle| y_{<n}^i, X^i, Z^{i-1}\right)$$

$$p\left(y_n^i \middle| y_{<n}^i, X^i, Z^{i-1}\right) = \mathrm{softmax}\left(W_o \widetilde{h_n^i}\right), \qquad \widetilde{h_n^i} = W_h[\boldsymbol{h}_n^i; \boldsymbol{c}_n^i; \boldsymbol{c}_n^{i-1}]$$



The knowledge of the CNMT

- The context-aware NMT tackles coreference resolution. [Tiedemann+, DiscoMT17, De-En], [Voita+, ACL18, En-Ru]
  - En-Ru MT: Pronouns in Russian are inflected by nouns in context sentences.
  - There is a possibility that multi-encoder-based CNMT can use the context information.

- [Bawden+, NAACL18] say ...
  - The **source-side context** is useful for CNMT.
  - The **target-side context** is not useful for CNMT.

  (context = previous sentence)

Hypothesis

- [Bawden+, NAACL18] only tried En-Fr translation.
  - Is this tendency in distant language pairs, such as En-Ja?

- Is the model consistent?
  - When the model uses <u>the source-side context</u>
  → the source-side sentence is incorporated into an <u>**encoder**</u>

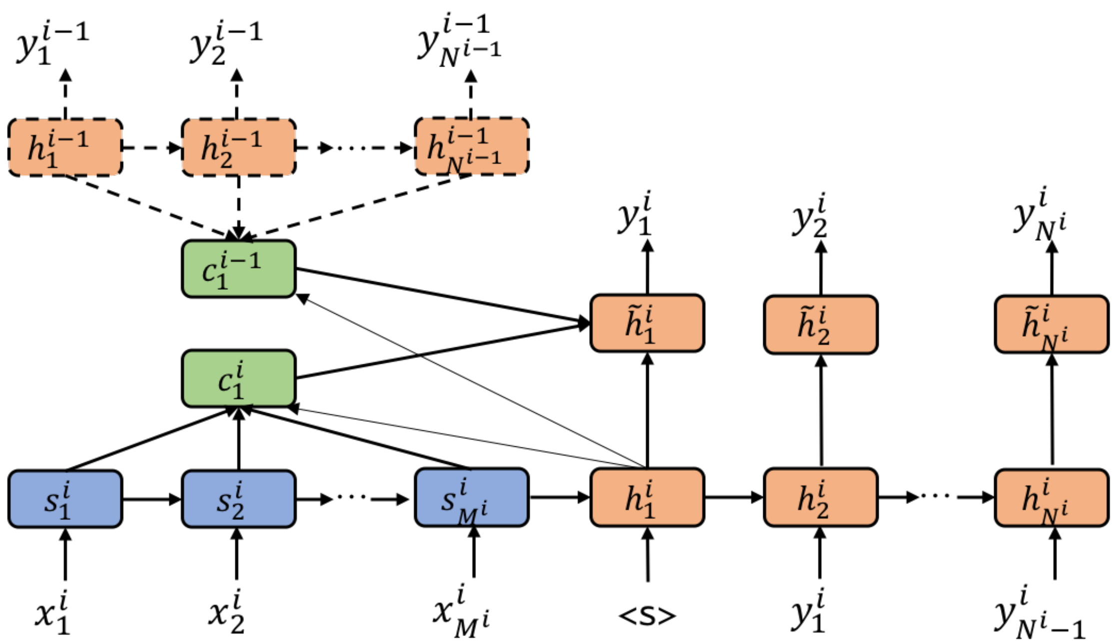
  - When the model uses <u>the target-side context</u>
  → the target-side sentence is also incorporated into an <u>**encoder**</u>

- In this research, we hypothesized that
  the target-side sentence should be incorporated into a <u>decoder.</u>

Proposed method: shared model

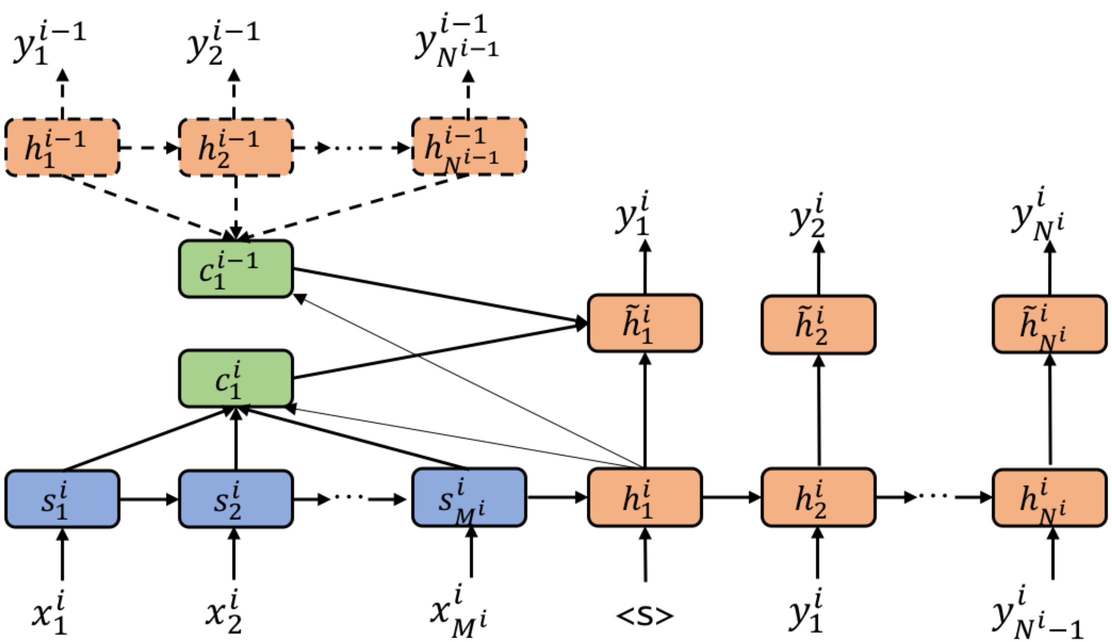
- The shared model saves hidden states of the decoder.
- When translating the current sentence, this model uses the saved states as an output of the context encoder.

- This model doesn't require many additional parameters and much computational times.



Shared source and shared target

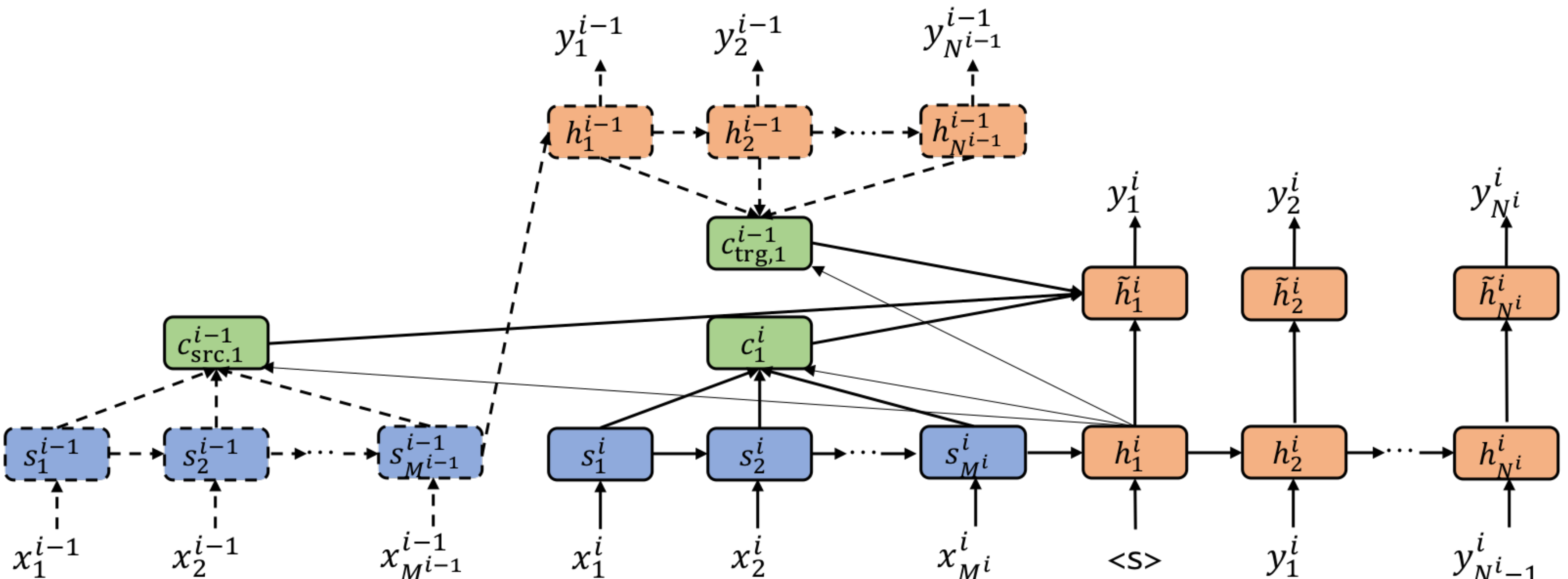
- Shared source: using the source-side context.
- Shared target: using the target-side context.
  - Saving the hidden states of <u>decoder</u> as a context.
  - <u>**The target-side context can be incorporated into a decoder instead of an encoder.**</u>



Shared mix

- If we use the context of both sides, what will happen?
  → Shared Mix model

- Saving the both sides of hidden states, and using them.
- For keeping the parameter size, attention is $c_{src}^{i-1} + c_{trg}^{i-1}$.



Related works of context-aware NMT

- Hierarchical Encoder [Wang+, EMNLP17, Zh-En task]
  - First encoder: word embedding → sentence embedding
  - Second encoder: sentence embedding → document embedding

  - This model can use the context of several sentences.

- Cache [Tu+, TACL18, Zh-En task]
  - Saving generated words and hidden states in a cache.
  - Calculating the attention with the cache.

  - They found, “more than 5 contexts is not needed.”

→ Both models need a huge neural architecture.

Experiment

Datasets

- TED Corpus (Subtitles on TED Talks)
  - 6 language pairs: De-En, En-De, Zh-En, En-Zh, Ja-En, En-Ja
- Recipe Corpus (User-posted recipes on popular website in Japan)
  - 2 language pairs: Ja-En, En-Ja

	Language Family	Word order	Context
En-De	Same	Similar (SVO - SOV (V2))	Low context
En-Zh	Different	Same (SVO)	Low context
En-Ja	Different	SVO - SOV	High context (Ja)

Corpus	Tokenizer	Train	Dev	Test
TED De-En	Moses Tokenizer	203,998	888	1,305
TED Zh-En	Jieba / Moses Tokenizer	226,196	879	1,297
TED Ja-En	MeCab / Moses Tokenizer	194,170	871	1,285
Recipe Ja-En	MeCab / Moses Tokenizer	108,990	3,303	2,804

Experimental Setup

- Baseline: Global Attention NMT [Luong+, EMNLP15]
  - Hyper-parameters are as follows.
- BPE (the algorithm of subword tokenization [Sennrich+, ACL16]) was applied to the vocabulary of each language.
- Metrics: BLEU [Papineni+, ACL02]

Hyper-parameter	setup
Dimension of embedding	512
Dimension of hidden state	512
Vocabulary (subword level)	TED: 32,000, Recipe: 8,000
Minibatch size	128 documents
Encoder / Decoder	2-layer Bi-LSTM / 2-layer Uni-LSTM
Beam size	5
Optimization	AdaGrad (Initial learning rate: 0.01)

Experimental Setup (for proposed method)

- Pretrained with baseline models, except for the context encoder.
- When the model translates the first sentence of document, the context attention $c^0 = \mathbf{0}$
- For separated models
  - The context encoder is initialized with random values
  - The reference is used as an input of the context encoder during training of the separated target model.
- For shared models
  - The saved hidden states are used during training and test.
- Each experiment is executed three times.
  - We show the average, the SD, and statistical significance calculated by bootstrap resampling.

Result

- The **shared target** improves the performance in all pairs.
- Improvement of the **separated target** is less compared to the baseline.
- The target-side context should be introduced to a decoder.

Experiment	Baseline	Separated		Shared		
		Source	Target	Source	Target	Mix
TED De–En	26.55	26.29 ± .37	26.52 ± .12	*27.20 ± .11	* **27.34** ± .11	27.18 ± .21
TED En–De	21.26	21.04 ± .64	20.77 ± .10	21.63 ± .27	**21.83** ± .30	21.50 ± .29
TED Zh–En	12.54	12.52 ± .33	12.63 ± .24	*13.36 ± .41	* **13.52** ± .10	*13.23 ± .09
TED En–Zh	8.97	8.94 ± .11	8.71 ± .06	9.45 ± .22	* **9.58** ± .13	9.42 ± .19
TED Ja–En	5.84	*6.64 ± .26	*6.37 ± .12	*6.95 ± .07	* **6.96** ± .18	*6.81 ± .16
TED En–Ja	8.40	8.58 ± .12	8.26 ± .00	8.51 ± .31	8.59 ± .08	**8.66** ± .14
Recipe Ja–En	25.34	*26.51 ± .09	*26.69 ± .15	*26.90 ± .17	* **26.92** ± .10	*26.78 ± .11
Recipe En–Ja	20.81	*21.87 ± .12	*21.45 ± .14	* **22.02** ± .20	*21.97 ± .09	*21.81 ± .15

Result

- The **<u>shared source</u>** also improves the performances.
- Is weight sharing more efficient than learning contexts?
  - Weight sharing between the layers keeps performance. [Dabre+, AAAI19]
  - It can be seen as an instance of multitask learning between translation and skip-thought.

Experiment	Baseline	Separated		Shared		
		Source	Target	Source	Target	Mix
TED De–En	26.55	26.29 ± .37	26.52 ± .12	*27.20 ± .11	***27.34** ± .11	27.18 ± .21
TED En–De	21.26	21.04 ± .64	20.77 ± .10	21.63 ± .27	**21.83** ± .30	21.50 ± .29
TED Zh–En	12.54	12.52 ± .33	12.63 ± .24	*13.36 ± .41	***13.52** ± .10	*13.23 ± .09
TED En–Zh	8.97	8.94 ± .11	8.71 ± .06	9.45 ± .22	***9.58** ± .13	9.42 ± .19
TED Ja–En	5.84	*6.64 ± .26	*6.37 ± .12	*6.95 ± .07	***6.96** ± .18	*6.81 ± .16
TED En–Ja	8.40	8.58 ± .12	8.26 ± .00	8.51 ± .31	8.59 ± .08	**8.66** ± .14
Recipe Ja–En	25.34	*26.51 ± .09	*26.69 ± .15	*26.90 ± .17	***26.92** ± .10	*26.78 ± .11
Recipe En–Ja	20.81	*21.87 ± .12	*21.45 ± .14	***22.02** ± .20	*21.97 ± .09	*21.81 ± .15

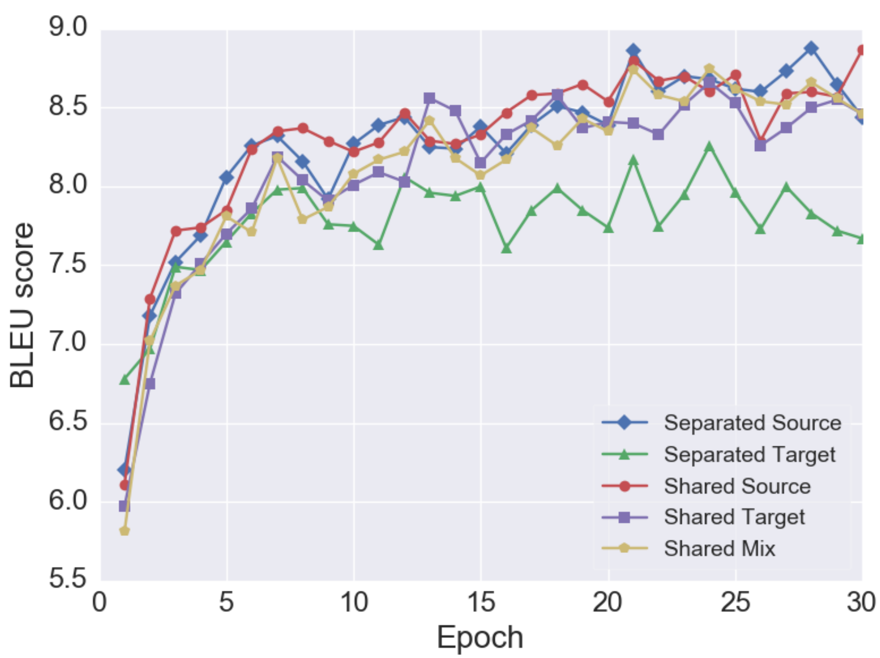
Result

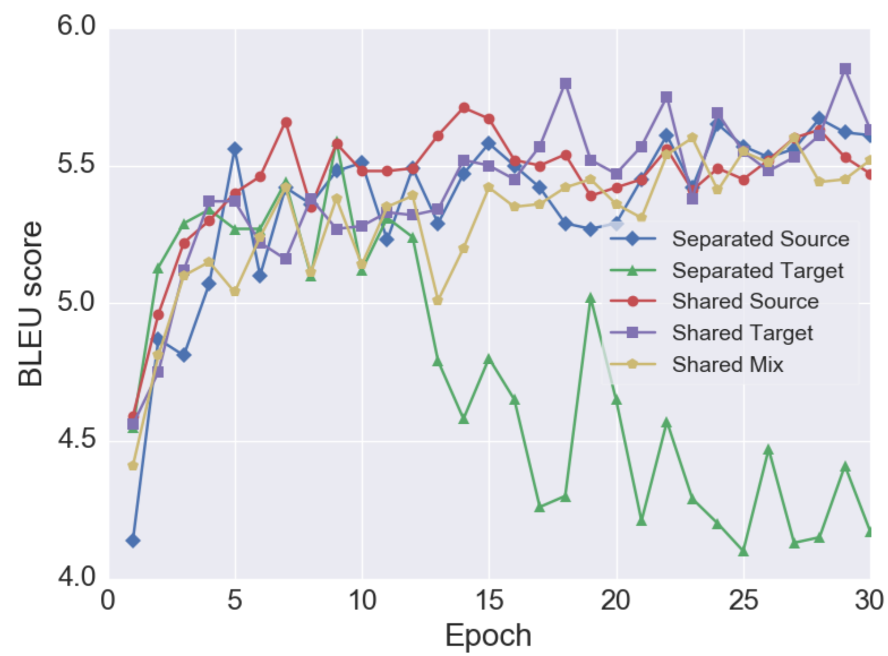
- In Ja-En / En-Ja results, both sides of context are equally important
- In other cases, target-side context is more important.
  - Does the word order have an important role in CNMT?
  - High context language (Japanese) has unique characteristics?

Experiment	Baseline	Separated		Shared		
		Source	Target	Source	Target	Mix
TED De–En	26.55	26.29 ± .37	26.52 ± .12	*27.20 ± .11	***27.34** ± .11	27.18 ± .21
TED En–De	21.26	21.04 ± .64	20.77 ± .10	21.63 ± .27	**21.83** ± .30	21.50 ± .29
TED Zh–En	12.54	12.52 ± .33	12.63 ± .24	*13.36 ± .41	***13.52** ± .10	*13.23 ± .09
TED En–Zh	8.97	8.94 ± .11	8.71 ± .06	9.45 ± .22	***9.58** ± .13	9.42 ± .19
TED Ja–En	5.84	*6.64 ± .26	*6.37 ± .12	*6.95 ± .07	***6.96** ± .18	*6.81 ± .16
TED En–Ja	8.40	8.58 ± .12	8.26 ± .00	8.51 ± .31	8.59 ± .08	**8.66** ± .14
Recipe Ja–En	25.34	*26.51 ± .09	*26.69 ± .15	*26.90 ± .17	***26.92** ± .10	*26.78 ± .11
Recipe En–Ja	20.81	*21.87 ± .12	*21.45 ± .14	***22.02** ± .20	*21.97 ± .09	*21.81 ± .15

Convergence of training

- Graph: the BLEU scores on development sets
- The <u>separated target</u> is unstable.
  - This is due to the exposure bias between a decoder and a context encoder
  - The shared target is stable because of no exposure biases.





	Sentence (the upper sentences represent context sentence.)
Input	わかめはよく洗って塩を落とし、10分ほど水に浸けておいてからざく切りにする。長ねぎは小口切りにする。
	熱した鍋にごま油をひき、わかめと長ねぎを入れて30秒ほど軽く炒める。
Reference	Wash the wakame well to remove the salt, put into a bowl of water for 10 minutes and drain. Cut into large pieces. Slice the Japanese leek.
	Heat a pan and pour the sesame oil. Stir fry the wakame and leek for 30 seconds.
Baseline	Wash the wakame seaweed well and remove the salt. Soak in water for 10 minutes, then roughly chop. Cut the Japanese leek into small pieces.
	Heat sesame oil in a heated pot, add the wakame and leek, and lightly sauté for about 30 seconds.
Separated Target	Wash the wakame well, soak in water for about 10 minutes. Cut into small pieces. Cut the Japanese leek into small pieces.
	Heat the sesame oil in a frying pan, add the wakame and leek, and stir-fry for about 30 seconds.

	Sentence (the upper sentences represent context sentence.)
Input	わかめはよく洗って塩を落とし、10分ほど水に浸けておいてからざく切りにする。長ねぎは小口切りにする。
	熱した鍋にごま油をひき、わかめと長ねぎを入れて30秒ほど軽く炒める。
Reference	Wash the wakame well to remove the salt, put into a bowl of water for 10 minutes and drain. Cut into large pieces. Slice the Japanese leek.
	Heat a pan and pour the sesame oil. Stir fry the wakame and leek for 30 seconds.
Baseline	Wash the wakame seaweed well and remove the salt. Soak in water for 10 minutes, then roughly chop. Cut the Japanese leek into small pieces.
	Heat sesame oil in a heated pot, add the wakame and leek, and lightly sauté for about 30 seconds.
Shared Target	Wash the wakame well, remove the salt, soak in water for about 10 minutes, then roughly chop. Chop the Japanese leek into small pieces.
	Heat sesame oil in a heated pan, add the wakame and Japanese leek, and lightly stir-fry for about 30 seconds.

Conclusion

- We reported how CNMT effectively employs the target-side context.
- The shared model can incorporate the target-side context into a decoder instead of an encoder.
- The shared model achieves the high performance, even though it does not need additional costs.
- We found that the importance of context is different between the language pairs.
- Our code is available: https://github.com/hargon24/Context_aware_NMT