HANOIT: Enhancing Context-aware Translation via Selective Context

Jian Yang[1], Yuwei Yin[2], Shuming Ma[3], Liqun Yang[1(✉)], Hongcheng Guo[1], Haoyang Huang[3], Dongdong Zhang[3], Yutao Zeng[1], Zhoujun Li[1], and Furu Wei[2]

[1] State Key Lab of Software Development Environment, Beihang University, Beijing, China
{jiaya,lqyang,hongchengguo,zengyutao,lizj}@buaa.edu.cn
[2] The University of Hong Kong, Hong Kong, China
yuweiyin@hku.hk
[3] Microsoft Research Asia, Beijing, China
{shumma,haohua,dozhang,fuwei}@microsoft.com

Abstract. Context-aware neural machine translation aims to use the document-level context to improve translation quality. However, not all words in the context are helpful. The irrelevant or trivial words may bring some noise and distract the model from learning the relationship between the current sentence and auxiliary context. To mitigate this problem, we propose a novel end-to-end encoder-decoder model with a layer-wise selection mechanism to sift and refine the long document context. To verify the effectiveness of our method, extensive experiments and extra quantitative analysis are conducted on four document-level machine translation benchmarks. The experimental results demonstrate that our model significantly outperforms previous models on all datasets via the soft selection mechanism.

Keywords: Neural Machine Translation · Context-aware Translation · Soft Selection Mechanism

1 Introduction

Recently, neural machine translation (NMT) based on the encoder-decoder framework has achieved state-of-the-art performance on the sentence-level translation [2,5,6,23,24,26,31–33]. However, the sentence-level translation solely considers single isolated sentence in the document and ignores the semantic knowledge and relationship among them, causing difficulty in dealing with the discourse phenomenon such as lexis, ellipsis, and lexical cohesion [27,30].

To model the document-level context, there are two main context-aware neural machine translation schemes. One approach introduces an additional context encoder to construct dual-encoder structure, which encodes the current source sentence and context sentences separately and then incorporates them via the gate mechanism [3,4,9,16,28,29]. The other one directly concatenates the current source sentence and context sentences as a whole input to the standard Transformer architecture, though the input sequence might be quite long

X. Wang et al. (Eds.): DASFAA 2023, LNCS 13945, pp. 471–486, 2023.
https://doi.org/10.1007/978-3-031-30675-4_34

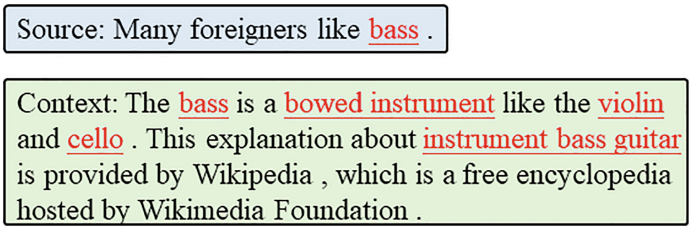

Fig. 1. An example of the source and the context sentence. Above is a source sentence to be translated, and below is its context in the same document. The underlined words are useful to disambiguate the source sentence, while the rest is less important.

[1,3,20,25]. The previous works [1,13] conclude that the Transformer model has the capability to capture long-range dependencies, where the self-attention mechanism enables the simple concatenation method to have competitive performance with multi-encoder approaches.

Most aforementioned previous methods use the whole context sentences and assume that all words in the context have a positive effect on the final translation. Despite the benefits of part of the context, not all context words are useful to the current translation. In Fig. 1, the underlined words provide supplementary information for disambiguation, while the others are less important. The irrelevant words may bring some noise and redundant content, increasing the difficulty for the model to learn the relationship between the context and the translation. Therefore, these useless words should be discarded so that the model can focus on the relevant information of the current sentence.

In this work, we propose an end-to-end model to translate the source document based on layer-wise context selection over encoder. In our model, the context is concatenated with current source sentence as external knowledge to be fed into the unified self-attention, where they are precisely selected among multiple layers to gradually discard useless information. The criteria on context selection is based on context-to-source attention score which are recursively calculated layer-by-layer. Ultimately, the context on the top layer is expected to be the most useful knowledge to help current source sentence translation. The architecture of our model looks like a Tower of **Hanoi** over the **T**ransformer structure (HANOIT). Our proposed model captures all context words at the bottom layer and focuses more on the essential parts at the top layer via the soft selection mechanism.

To verify the effectiveness of our method, we conduct main experiments and quantitative analysis on four popular benchmarks, including IWSLT-2017, NC-2016, WMT-2019, and Europarl datasets. Experimental results demonstrate that our method significantly outperforms previous baselines on these four popular benchmarks and can be further enhanced by the sequence-to-sequence pretrained model, such as BART [12]. Analytic experiments and attention visualization illustrate our proposed selection mechanism for avoiding the negative interference introduced by noisy context words and focusing more on advantageous context pieces.

2 Our Approach

In this section, we will describe the architecture of our HanoiT, and apply HanoiT to context-aware machine translation.

2.1 Problem Statement

Formally, let $X = \{x^{(1)}, .., x^{(k)}, .., x^{(K)}\}$ denote a source language document composed of K source sentences, and $Y = \{y^{(1)}, .., y^{(k)}, .., y^{(K)}\}$ is the corresponding target language document. $\{x^{(k)}, y^{(k)}\}$ forms a parallel sentence, where $x^{(k)}$ denotes the k^{th} source sentence and $y^{(k)}$ is the translation of $x^{(k)}$. $X_{<k} = \{x^{(1)}, .., x^{(k-1)}\}$ denotes the historical context of $x^{(k)}$ and $X_{>k} = \{x^{(k+1)}, .., x^{(K)}\}$ represents the future context. Given the current source sentence $x^{(k)}$, the historical context $X_{<k}$, and the future context $X_{>k}$, the translation probability is calculated by:

$$P(y^{(k)}|X;\theta) = \prod_{i=1}^{N} P(y_i^{(k)}|X, y_{<i}^{(k)};\theta) \tag{1}$$

where $y_i^{(k)}$ is the i^{th} word of the k^{th} target sentence and $y_{<i}^{(k)}$ are the previously generated words of the target sentence $y^{(k)}$ before i^{th} position. $y^{(k)}$ has N words. In this work, we use one previous and one next sentence as the context.

2.2 HanoiT

Figure 2 shows the overall structure of our HanoiT model. At the bottom of the encoder, it models the concatenation of the source sentence and the context with unified self-attention layers. At the top of the encoder, it gradually selects the context words according to the attention weights.

Embedding. We use the segment embedding to distinguish the current sentence, source, and target context sentences. In Fig. 2, we concatenate the current sentence and the source context as a whole. To model the positions of the different parts, we also reset the positions of the current source sentence and source context sentences. Therefore, the final embedding of the input words is the sum of the word embedding, position embedding, and segment embedding, which can be described as:

$$E = E_w + E_p + E_s \tag{2}$$

where E_w is the word embedding, E_s is the segment embedding from the learned parameter matrix, and E_p is the position embedding.

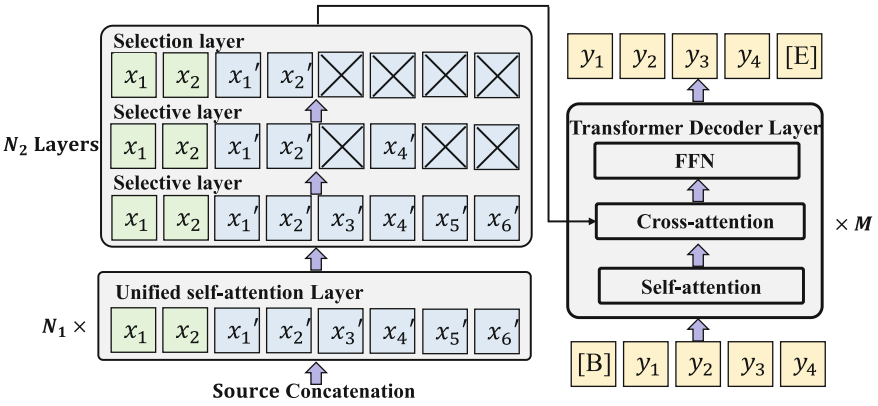

Fig. 2. Overview of our proposed model HANOIT. For simplicity, layer normalization and other components of the Transformer architecture are omitted in the picture. Cross symbols denote dropped words. (x_1, x_2) is the current source sentence and $(x'_1, x'_2, x'_3, x'_4, x'_5, x'_6)$ is the source context. N_1 and N_2 denote the number of unified self-attention layers and selection layers. $(x'_1, x'_2, x'_3, x'_4, x'_5, x'_6) \rightarrow (x'_1, x'_2, x'_4) \rightarrow (x'_1, x'_2)$ is the selective procedure, where important words are selected gradually by multiple selection layers.

Encoder. Since the inputs of context-aware neural machine translation are composed of several sentences, we build our model based on the multi-head attention to capture long-range dependencies and compute the representation of the document-level context. Our encoder consists of two groups of layers: unified self-attention layers and selection layers. The unified self-attention layers is to compute a joint representation of the source sentence and the context, while the selection layer is to select the context for the next layer.

Unified Self-attention Layer. Given the concatenation of the source sentence and the source context, we obtain the document representation $s^0 = \{s_1^0, .., s_p^0, .., , s_m^0\}$ after the embedding layer, where p is the length of $x^{(k)}$ and m is the length of source concatenation. Then, we feed the s^0 into N_1 unified self-attention layers to compute their representations.

$$s^l = \text{FFN}(\text{MultiHeadAttn}(s^{l-1}; \theta_{N_1})) \tag{3}$$

where the l is the number of the unified self-attention layer and $l \in [1, N_1]$.

Selection Layer. After N_1 unified self-attention layers, we get representations of source concatenation $s^{N_1} = \{s_1^{N_1}, .., s_p^{N_1}, .., s_m^{N_1}\}$, which can be used to select important context words. In the selection layer, we apply multi-head attention to s^{N_1}, and then average attention scores across different heads, which can be described as below:

$$a_{i,j} = \frac{1}{h} \sum_{1 \le i \le h} \text{MultiHeadAttn}(s^{N_1}) \tag{4}$$

where h is the number of attention heads. $a_{i,j}$ represents the average attention score between the i^{th} token and the j^{th} token.

Then, we calculate the average attention score between i^{th} word and other tokens in the source current sentence $x^{(k)}$:

$$a_{i,\neq i} = \frac{1}{p} \sum_{j \in [1,p], j \neq i} a_{i,j} \tag{5}$$

where $a_{i,\neq i}$ represents the average correlation between i^{th} word and the other words, and p is the number of tokens in the source sentence.

In order to decide which context words should be selected, we compute the correlation scores s between each context word and the whole source sentence. For the k^{th} context word, we count how many words in the current sentence have a higher attention score with it compared to the average attention score $a_{i,\neq i}$:

$$s_k = \sum_{i \in [p+1,m]} \delta_{a_{i,k} \geq a_{i,\neq i}} \tag{6}$$

where $\delta_{a_{i,k} \geq a_{i,\neq i}}$ equals 1 if $a_{i,k} \geq a_{i,\neq i}$ else 0, p is the number of tokens in the source sentence, and m is the total number of tokens in the concatenation of the source sentence and the source context.

Finally, we can select the context words with top correlation scores s_k. We use v_k to denote whether the k^{th} word is selected:

$$v_k = \delta_{s_k \geq q * p} \tag{7}$$

where $\delta_{s_k \geq q*p}$ equals to 1 if $s_k \geq q * p$ else 0, p is the number of tokens in the source sentence, and $q \leq 1$ is a hyper-parameter to control the percentage of the selective context. In this work, we set $q \in [0.1, 0.5]$ according to the performance in the validation set.

Decoder. The source selective concatenation $s^{N_2} = \{s_1^{N_2}, .., s_p^{N_2}, .., s_{m_1}^{N_2}\}$ is fed into the standard Transformer decoder to predict the final translation.

2.3 Bi-lingual Context Integration

Section 2.2 only considers the mono-lingual context, i.e. the source context. In practice, when translating a document, we can also obtain the target context by sentence-level translating the document before context-aware translation [27]. In this section, we extend our HanoiT model to integrate the bi-lingual context, i.e. the source context and the target context.

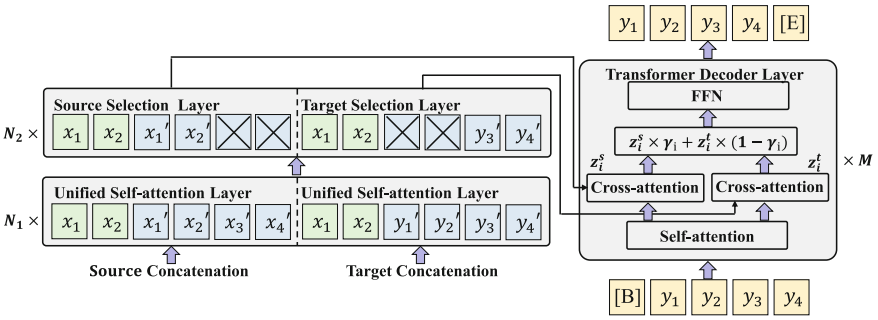

Fig. 3. Overview of the extended HANOIT to integrate the bilingual context. Cross symbols denote masked words. Source concatenation consists of the current source sentence (x_1, x_2) and source context (x'_1, x'_2, x'_3, x'_4). Target concatenation is composed of the source sentence (x_1, x_2) and the target context (y'_1, y'_2, y'_3, y'_4). Then the source and target selective concatenations are incorporated by the gate mechanism to predict the final translation.

Formally, let $X = \{x^{(1)}, .., x^{(k)}, .., x^{(K)}\}$ denote a source language document composed of K source sentences and $Y = \{y^{(1)}, .., y^{(k)}, .., y^{(K)}\}$ denotes the sentence-level translation of X. $X_{<k}$ is the historical source context and $X_{>k}$ is the future source context. Similarity, we denote historical target context $\{y^{(1)}, .., y^{(k-1)}\}$ as $Y_{<k}$ and future target context $\{y^{(k+1)}, .., y^{(K)}\}$ as $Y_{>k}$. We model the translation probability that is conditioned on the bi-lingual source context $X_{\neq k}$ and target context $Y_{\neq k}$ as:

$$P(y^{(k)}|X;\theta) = \prod_{i=1}^{N} P(y_i^{(k)}|X, Y_{\neq k}, y_{<i}^{(k)};\theta) \tag{8}$$

where $y_i^{(k)}$ is the i^{th} word of the k^{th} target sentence and $y_{<i}^{(k)}$ are the previously generated words of the target sentence $y^{(k)}$ before i^{th} position.

Encoder. As shown in Fig. 3, the current source sentence and the source context are merged as the source concatenation. Besides, the current source sentence and the target context are also merged as the target concatenation. Both concatenations are fed into unified self-attention and selection layers to compute representations of source concatenation s^{N_2} and target concatenation t^{N_2}.

Decoder. With the above encoder, we obtain the representations of the selective source concatenation $s^{N_2} = \{s_1^{N_2}, .., s_p^{N_2}, .., s_{m_1}^{N_2}\}$ and the selective target concatenation $t^{N_2} = \{t_1^{N_2}, .., t_p^{N_2}, .., t_{n_1}^{N_2}\}$, where m_1 and n_1 are lengths of selective source and target concatenation. Given both selective concatenations, we deploy the multi-head attention by two attention components. Using query, key, value parameters (W_s^Q, W_s^V, W_s^K), the decoder gets the hidden state z_i^s. Similarly, another hidden state z_i^t is generated by the additional attention component

with parameters (W_t^Q, W_t^V, W_t^K). Considering the previous insight [10] that the gate network is a good component for bi-lingual context setting, we employ the gate mechanism to incorporate the source and target context.

Gate Mechanism. Given the i^{th} hidden states z_i^s and z_i^t, the gate mechanism can be described as:

$$\gamma_i = c\sigma(W_s z_i^s + U_t z_i^t + b) \tag{9}$$

where W_s and U_t are parameters matrices and b is a bias. $c \in [0, 1]$ is a hyper-parameter to control range of the gate weight. $\sigma(\cdot)$ is the sigmoid function.

$$z_i = (1 - \gamma_i) z_i^s + \gamma_i z_i^t \tag{10}$$

where z_i is the i^{th} decoder final hidden state derived from the source context and target context.

2.4 Training

Given the mono-lingual context only, the training objective is a cross-entropy loss function on the top of Eq. 1. The objective $\mathcal{L}_m$ is written as:

$$\mathcal{L}_m = -\sum_{X, y^{(k)} \in D} \log P_\theta(y^{(k)}|X) \tag{11}$$

where θ are model parameters.

Considering the bi-lingual context, the training objective $\mathcal{L}_b$ is calculated as:

$$\mathcal{L}_b = -\sum_{X, y^{(k)}, Y_{\neq k} \in D} \log P_\theta(y^{(k)}|X, Y_{\neq K}) \tag{12}$$

where θ are model parameters.

The quality of the target context depends on the sentence-level translation model, which may bring additional errors. To reduce the possible harm by these errors and make the training stable, our model optimizes a combination of the mono-lingual objective $\mathcal{L}_m$ and the bi-lingual objective $\mathcal{L}_b$:

$$\mathcal{L}_{all} = \alpha \mathcal{L}_m + (1 - \alpha)\mathcal{L}_b \tag{13}$$

where α is a scaling factor to balance two objectives between $\mathcal{L}_m$ and $\mathcal{L}_b$. We find when the value of α equals 0.5, our model gets the optimal performance by balancing two objectives. We adopt Eq. 11 to train the model with mono-lingual context, and Eq. 13 to train the model with bi-lingual context.

3 Experiments

To prove the efficiency of our method, we conduct experiments on four public benchmarks.

Table 1. Sentence-level evaluation results on four tasks with BLEU% metric using the source context. Bold numbers denote the best BLEU points. RNN and Transformer are context-agnostic baselines and others are context-aware baselines. The results with the symbol "†" are directly reported from the previous work. BLEU points with the symbol "*" are re-implemented by ourselves. "‡" denotes our proposed method.

Mono-lingual Context	IWSLT-2017	NC-2016	Europarl	WMT-2019
RNN [2]	19.24†	16.51†	26.26†	–
Transformer [26]	23.28†	22.78†	28.72†	–
Transformer (our re-implementation)	24.52*	24.45*	29.98*	38.02*
ECT [25]	24.32*	24.40*	30.08*	38.14*
Dual Encoder [9]	24.14*	24.36*	30.12*	38.12*
DCL [41]	24.00†	23.08†	29.32†	–
HAN [30]	24.58†	25.03†	28.60†	–
Transformer + QCN [40]	24.41†	22.22†	29.48†	–
SAN [16]	24.55†	24.78†	29.75†	–
Flat Transformer [13]	24.87†	23.55†	30.09†	38.34*
HanoiT (our method)	**24.94**‡	**25.22**‡	**30.49**‡	**38.52**‡

3.1 Datasets

To evaluate our method, we use the same dataset as previous work, including IWSLT-2017, NC-2016, Europarl, and WMT-2019 En-De translation [16].

IWSLT-2017. This corpus is from IWSLT-2017 MT track and contains transcripts of TED talks aligned at the sentence level.

NC-2016. NC-2016 dataset is from Commentary v9 corpus. Newstest2015 and newstest2016 are used as the valid and the test set.

Europarl. The dataset from Europarl v7 is split into training, valid and test sets according to the previous work [16]. Europarl is extracted from the European Parliament website.

WMT-2019. The WMT-2019 dataset comes from the WMT-2019 news translation shared task for English-German. Newstest2016, newstest2017, and newstest2018 are concatenated as the valid set. Newstest2019 is used as the test set.[1]

3.2 Implementation Details

Considering the model performance and computation cost, we use one previous and one next sentence as the source and target context for all our experiments.

[1] https://www.statmt.org/wmt19/.

Table 2. Sentence-level evaluation results on four tasks with BLEU% metric using the bi-lingual context. Bold numbers represent the best BLEU points. The results with the symbol "†" are directly reported from the previous work. BLEU points with the symbol "*" are re-implemented by ourselves. "‡" represents our proposed method.

Bi-lingual Context	IWSLT-2017	NC-2016	Europarl	WMT-2019
ECT [25]	24.38*	24.55*	30.24*	38.16*
Dual Encoder [9]	24.26*	24.46*	30.25*	38.24*
DCL [41]	23.82†	22.78†	29.35†	–
HAN [30]	24.39†	24.38†	29.58†	–
CADec [27]	24.45*	24.30*	29.88*	–
SAN [16]	24.62†	24.36†	29.80†	–
HanoiT (our method)	**25.04**‡	**25.28**‡	**30.89**‡	**38.55**‡

The evaluation metric is case-sensitive tokenized BLEU [18]. For different benchmarks, we adapt the batch size, the beam size, the length penalty, the number of unified self-attention layers N_1, and the number of selection layers N_2 to get better performance. For all experiments, we use a dropout of 0.1 and cross-entropy loss with a smoothing rate of 0.1 for sentence-level and context-aware baselines except notification. All sentences are tokenized with Moses [11] and encoded by BPE [21] with a shared vocabulary of 40K symbols. The batch size is limited to 2048 target tokens by default. For the **IWSLT-2017** dataset, we deploy the small setting of the Transformer model, which has 6 layers with 512 embedding units, 1024 feedforward units, 4 attention heads, a dropout of 0.3, a l_2 weight decay of 1e-4. For the **NC-2016** dataset, we use the base setting of Transformer [26], in which both the encoder and the decoder have 6 layers, with the embedding size of 512, feedforward size of 2048, and 8 attention heads. We set both dropout and attention dropout as 0.2 for our method. For the **Europarl** and the **WMT-2019** dataset, the base setting of the Transformer model with 4000 warming-up steps is used.

3.3 Baselines

For the mono-lingual and the bi-lingual context setting, we compare our method with other baselines.

Mono-lingual Context: **RNN** [2] and **Transformer** [26] are backbone models. **ECT** [25] simply concatenates the source sentence and context into the standard Transformer model. Besides, **Dual Encoder** [9] uses two encoders to incorporate the source sentence and context sentences to predict the translation. Moreover, **DCL** [41] incorporates context hidden states into both the source encoder and target decoder. **Flat Transformer** [13] focus on the current self-attention at the top. Furthermore, **HAN** [30] and **SAN** [16] introduce the hierarchical and selection attention mechanism. **QCN** [40] is a query-guided capsule networks.

Table 3. Sentence-level evaluation results on four benchmarks with BLEU% metric under the mono-lingual context setting. The architecture $N_1 + N_2$ represents our HANOIT consists of N_1 unified self-attention layers and N_2 selection layers. The architecture ($N_1 = 6$, $N_2 = 0$) only uses six unified self-attention layers with the segment embedding, which select all context words to generate the final translation.

Architecture	IWSLT-2017	NC-2016	Europarl	WMT-2019	Average
6 + 0	24.48	24.64	30.22	38.22	29.39
5 + 1	24.32	24.95	30.52	38.46	29.74
4 + 2	24.55	24.52	30.72	38.37	29.54
3 + 3	24.64	24.88	30.65	38.12	29.58
2 + 4	24.74	24.60	30.62	**38.62**	29.65
1 + 5	**24.94**	**25.22**	**30.49**	38.40	**29.85**
0 + 6	24.56	24.75	30.66	37.98	29.49

Bi-lingual Context: **CADec** [27] is composed of identical multi-head attention layers, of which the decoder has two multi-head encoder-decoder attention with encoder outputs and first-pass decoder outputs. Also, **Dual Encoder**, **ECT**, **DCL**, **HAN** and **SAN** can also use the bi-lingual context to improve the performance.

3.4 Main Results

Mono-lingual Context. We present the results of our proposed method, sentence-level baselines, and other context-aware baselines in Table 1, which all only use the mono-lingual source context. The context-aware baselines include ECT, Dual Encoder, DCL, HAN, SAN, and Flat Transformer. The sentence-level Transformer model gets 24.52, 24.55, 29.98, and 38.02 BLEU points on four benchmarks. Compared to this strong baseline, our model also significantly gains an improvement of +0.42, +0.77, +0.81, and +0.51 BLEU points respectively on four benchmarks. Furthermore, our method outperforms SAN by +0.39, +0.44, +0.74 BLEU points on IWSLT-2017, NC-2016, and Europarl datasets. We also observe that most context-aware models gain better performance than the sentence-level model Transformer, especially on IWSLT-2017, NC-2016, and Europarl datasets. We conjecture these three datasets are suitable for evaluating context-aware models, where the current sentence needs to learn longer dependencies.

Bi-lingual Context. Under the bi-lingual context setting, our method also outperforms other baselines, including Dual Encoder, ECT, DCL HAN, CADec,

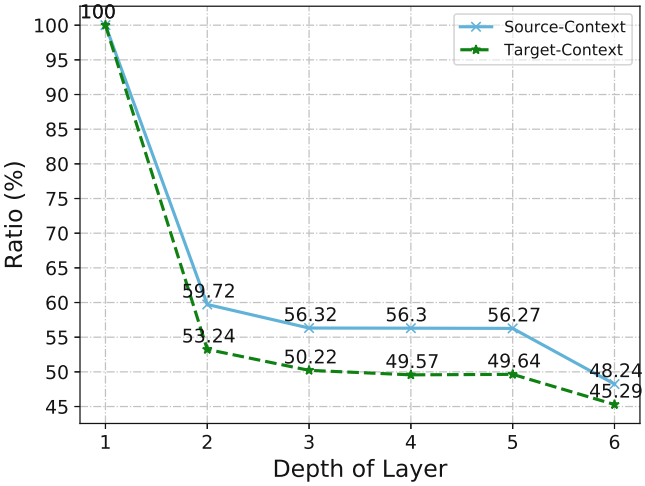

Fig. 4. Ratio of the source and target context selected words on the NC-2016 dataset. Our model selects useful words from both the source and the target context gradually layer by layer. Therefore, the number of context words reduces as the depth of selection layer increasing.

and SAN. HanoiT can achieve improvements of +1.02, +1.33, +0.91, +0.53 BLEU points than the sentence-level Transformer baseline. It proves that HanoiT also can be compatible with the target context to select useful words. Besides, HanoiT can significantly outperform the related baseline SAN model by +0.49, +0.50, +1.14 BLEU points, achieving better performance on three benchmarks. We also observe that the bi-lingual context provides marginal improvements over the mono-lingual context. According to these results, we infer that whether the context-aware model benefits from the bi-lingual context setting is dependent on the specific dataset.

4 Analysis

Attention Visualization. Our model encodes the concatenation of the source words and all context words by the unified attention layers at the bottom layers. As shown in Fig. 5(a), the model focuses on the source sentence "What do you do when you have a headache ?" and all context words "You swallow an aspir@@ in ." using the self-attention mechanism, which ensures that all context words can provide the external guidance and implicitly contribute to the translation. The context words with higher attention weights tend to be selected. In Fig. 5(b), the model only focuses on the source sentence and selected context word "aspir@@". The source word "headache" has a correlation with the context word "aspir@@". In this way, our method pays more attention to the current sentences and the selected words, while the other context words also provide the supplemental semantics for the current sentences at the bottom.

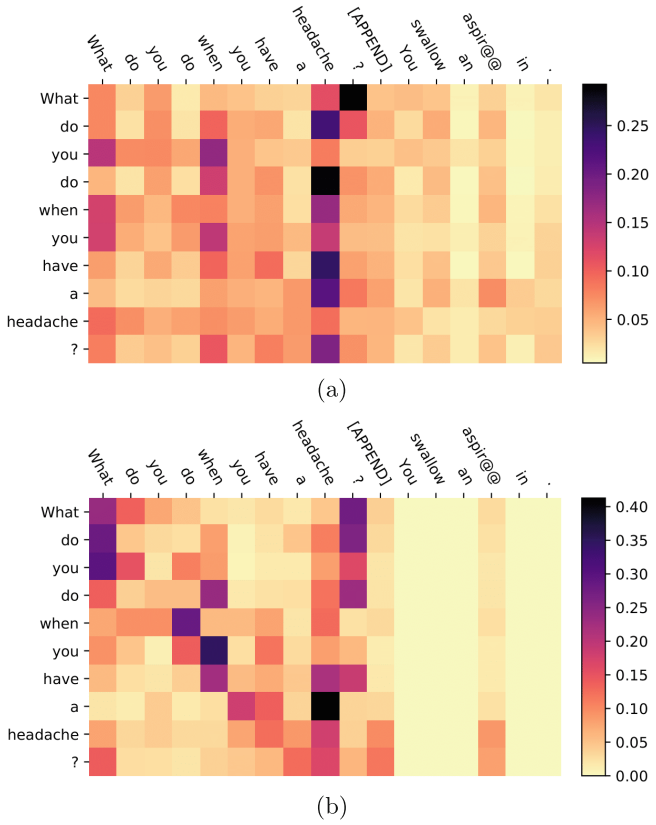

Fig. 5. Attention visualization of the encoder self-attention weights of the bottom unified attention layer (a) and top layer after the selection operation (b).

Number of Selection Layers. To better understand the impact of the selection layers for the translation performance, we tune different numbers of concatenation self-attention layers (N_1 layers) and selection layers (N_2 layers) to get the better performance under the mono-lingual setting. For the fair comparison, we keep the $N_1 + N_2 = 6$, which equals the number of the base setting Transformer layers. As shown in Table 3, we find that the architecture "$N_1 = 1$, $N_2 = 5$" gains the best average performance on four benchmarks. Besides, stacking too many selection layers also leads to worse performance, which may be caused by wrongly discarding too many context words. In summary, our proposed model uses all context words by unified self-attention layers, and focuses those important context words at the top of encoder blocks.

Ratio of Selected Words. We investigate how many words in the context are selected on the NC-2016 dataset. Figure 4 shows that the first selection layer reserves only 60% words from the context. After multiple selection layers, the

Table 4. Results of our method with the offline (1 previous + 1 next) and the online (1 previous) setting.

Context	IWSLT-2017	NC-2016	Avg.
Online (1 previous)	24.62	24.78	24.70
cre Offline (1 previous + 1 next)	**24.94**	**25.12**	25.07

ratio gradually reduces to 48.24%. Another obvious phenomenon is that the ratio of the target context is less than that of the source context. An intuitive explanation is that we use the source sentence words to select source or target context words, where source-source attention has a higher score compared with the source-target attention. Representations of the same language have a closer relationship than those of different languages on average [19].

Online vs. Offline Setting. Table 4 lists results of our method with different context settings. "1 previous" denotes the online setting where the context only includes one previous sentence. "1 previous + 1 next" denotes the offline setting where the context includes one previous and one next sentence. From the table, we can find that our proposed method has the similar performance with online and offline settings on the IWSLT-2017 and NC-2016 datasets. We also try the longer context including "2 previous + 2 next" and "3 previous + 3 next", but find no significant improvement.

Mono-lingual vs. Bi-lingual Context. For the source and target mono-lingual context setting, our method gets 24.94 and 24.98 BLEU points on the IWSLT-2017 dataset. Furthermore, we conduct experiments with the bi-lingual context sentences and get 25.04 BLEU points, where the target context sentences are the translation of the source context sentences. The bi-lingual context setting of our method has limited improvement over the mono-lingual context setting. The reason is that the target-side context shares similar information to the source-side, which also has been found by the previous work [16].

Leveraging Pre-trained Model. Since the parameters of our model are the same as the standard Transformer, our model can be initialized with the pre-trained model to enhance our method. The pre-trained model BART-large [12] is used for initialization under the mono-lingual context setting. We extract 12 bottom layers of the BART encoder and 6 bottom layers of the BART decoder to initialize our model. On IWSLT-2017 dataset, our model gains +1.91 BLEU improvement (24.94 → 26.85) with pre-trained model BART.

5 Related Work

Sentence-level Machine Translation. Sentence-level neural machine translation has developed immensely in the past few years, from RNN-based [2,6,24,32,33,

36], CNN-based [5], to the self-attention-based architecture [23,26,31,34,35,37–39]. However, these models always performed in a sentence-by-sentence manner, ignoring the long-distance dependencies. The past or future context can be important when it refers to using discourse features to translate the source sentence to the target sentence.

Context-aware Machine Translation. Context-aware machine translation aims to incorporate the source or target context to help translation. Previous works [3,7,9,15,22,27–29] have proven the importance of context in capturing different types of discourse phenomena such as Deixis, Ellipsis, and Lexical Cohesion. Others [3,9,16,28,29] explore the dual encoders and concatenation-based context-aware models.

Recently, a promising line of research to improve the performance of the context-aware NMT is to select useful words of the whole context, which can be used to enhance the positive use of context [8,17,42]. Other researchers propose selective attention mechanism by introducing sparsemax function [14,16].

6 Conclusion

In this work, we explore the solution to select useful words from the context. We propose a novel model called HANOIT, consisting of unified self-attention layers and selection layers. The experiments on both mono-lingual and bi-lingual context settings further prove the effectiveness of our method Experimental results demonstrate that our proposed method can select useful words to yield better performance.

References

1. Agrawal, R.R., Turchi, M., Negri, M.: Contextual handling in neural machine translation: look behind, ahead and on both sides. In: EAMT 2018, pp. 11–20 (2018)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR 2015 (2015)
3. Bawden, R., Sennrich, R., Birch, A., Haddow, B.: Evaluating discourse phenomena in neural machine translation. In: NAACL 2018, pp. 1304–1313 (2018)
4. Chen, L., et al.: Improving context-aware neural machine translation with source-side monolingual documents. In: IJCAI 2021, pp. 3794–3800. https://www.ijcai.org/ (2021)
5. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.N.: Convolutional sequence to sequence learning. In: ICML 2017, pp. 1243–1252 (2017)
6. Geng, X., Feng, X., Qin, B., Liu, T.: Adaptive multi-pass decoder for neural machine translation. In: EMNLP 2018, pp. 523–532 (2018)
7. Gonzales, A.R., Mascarell, L., Sennrich, R.: Improving word sense disambiguation in neural machine translation with sense embeddings. In: WMT 2017, pp. 11–19 (2017)
8. Jean, S., Cho, K.: Context-aware learning for neural machine translation. CoRR abs/1903.04715 (2019)

9. Jean, S., Lauly, S., Firat, O., Cho, K.: Does neural machine translation benefit from larger context? CoRR abs/1704.05135 (2017)
10. Junczys-Dowmunt, M.: Microsoft translator at WMT 2019: towards large-scale document-level neural machine translation. In: WMT 2019, pp. 225–233 (2019)
11. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: ACL 2007, pp. 177–180 (2007)
12. Lewis, M., et al.: BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: ACL 2020, pp. 7871–7880 (2020)
13. Ma, S., Zhang, D., Zhou, M.: A simple and effective unified encoder for document-level machine translation. In: ACL 2020 (2020)
14. Martins, A.F.T., Astudillo, R.F.: From softmax to sparsemax: a sparse model of attention and multi-label classification. In: ICML 2016, pp. 1614–1623 (2016)
15. Maruf, S., Haffari, G.: Document context neural machine translation with memory networks. In: ACL 2018, pp. 1275–1284 (2018)
16. Maruf, S., Martins, A.F.T., Haffari, G.: Selective attention for context-aware neural machine translation. In: NAACL 2019, pp. 3092–3102 (2019)
17. Maruf, S., Saleh, F., Haffari, G.: A survey on document-level machine translation: methods and evaluation. CoRR abs/1912.08494 (2019)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: ACL 2002, pp. 311–318 (2002)
19. Qin, L., Ni, M., Zhang, Y., Che, W.: CoSDA-ML: multi-lingual code-switching data augmentation for zero-shot cross-lingual NLP. In: IJCAI 2020, pp. 3853–3860 (2020)
20. Scherrer, Y., Tiedemann, J., Loáiciga, S.: Analysing concatenation approaches to document-level NMT in two different domains. In: EMNLP 2019, pp. 51–61 (2019)
21. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: ACL 2016, pp. 1715–1725 (2016)
22. Smith, K.S., Aziz, W., Specia, L.: The trouble with machine translation coherence. In: EAMT 2016, pp. 178–189 (2016)
23. Stern, M., Chan, W., Kiros, J., Uszkoreit, J.: Insertion transformer: flexible sequence generation via insertion operations. In: ICML 2019, pp. 5976–5985 (2019)
24. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NIPS 2014, pp. 3104–3112 (2014)
25. Tiedemann, J., Scherrer, Y.: Neural machine translation with extended context. In: EMNLP 2017, pp. 82–92 (2017)
26. Vaswani, A., et al.: Attention is all you need. In: NIPS 2017, pp. 5998–6008 (2017)
27. Voita, E., Sennrich, R., Titov, I.: When a good translation is wrong in context: context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In: ACL 2019, pp. 1198–1212 (2019)
28. Voita, E., Serdyukov, P., Sennrich, R., Titov, I.: Context-aware neural machine translation learns anaphora resolution. In: ACL 2018, pp. 1264–1274 (2018)
29. Wang, L., Tu, Z., Way, A., Liu, Q.: Exploiting cross-sentence context for neural machine translation. In: EMNLP 2017, pp. 2826–2831 (2017)
30. Werlen, L.M., Ram, D., Pappas, N., Henderson, J.: Document-level neural machine translation with hierarchical attention networks. In: EMNLP 2018, pp. 2947–2954 (2018)
31. Wu, F., Fan, A., Baevski, A., Dauphin, Y.N., Auli, M.: Pay less attention with lightweight and dynamic convolutions. In: ICLR 2019 (2019)
32. Wu, Y., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144 (2016)

33. Xia, Y., et al.: Deliberation networks: sequence generation beyond one-pass decoding. In: NIPS 2017, pp. 1784–1794 (2017)
34. Yang, J., et al.: GanLM: encoder-decoder pre-training with an auxiliary discriminator. CoRR **abs/2212.10218** (2022). https://doi.org/10.48550/arXiv.2212.10218
35. Yang, J., et al.: Multilingual machine translation systems from microsoft for WMT21 shared task. In: WMT@EMNLP 2021, pp. 446–455 (2021)
36. Yang, J., Ma, S., Zhang, D., Li, Z., Zhou, M.: Improving neural machine translation with soft template prediction. In: ACL 2020, pp. 5979–5989 (2020)
37. Yang, J., et al.: Learning to select relevant knowledge for neural machine translation. In: Wang, L., Feng, Y., Hong, Yu., He, R. (eds.) NLPCC 2021. LNCS (LNAI), vol. 13028, pp. 79–91. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88480-2_7
38. Yang, J., Yin, Y., Ma, S., Zhang, D., Li, Z., Wei, F.: High-resource language-specific training for multilingual neural machine translation. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, pp. 4461–4467. ijcai.org (2022). https://doi.org/10.24963/ijcai.2022/619
39. Yang, J., et al.: UM4: unified multilingual multiple teacher-student model for zero-resource neural machine translation. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, pp. 4454–4460. ijcai.org (2022). https://doi.org/10.24963/ijcai.2022/618
40. Yang, Z., Zhang, J., Meng, F., Gu, S., Feng, Y., Zhou, J.: Enhancing context modeling with a query-guided capsule network for document-level translation. In: EMNLP 2019, pp. 1527–1537 (2019)
41. Zhang, J., et al.: Improving the transformer translation model with document-level context. In: EMNLP 2018, pp. 533–542 (2018)
42. Zheng, Z., Huang, S., Sun, Z., Weng, R., Dai, X., Chen, J.: Learning to discriminate noises for incorporating external information in neural machine translation. CoRR abs/1810.10317 (2018)