Studying The Impact Of Document-level Context On Simultaneous Neural Machine Translation

Raj Dabre raj.dabre@nict.go.jp
National Institute of Information and Communications Technology

Aizhan Imankulova aizhan.imankulova@cogsmart-global.com
CogSmart

Masahiro Kaneko masahiro.kaneko@nlp.c.titech.ac.jp
Tokyo Institute of Technology

Abstract

In a real-time simultaneous translation setting, neural machine translation (NMT) models start generating target language tokens from incomplete source language sentences, making them harder to translate, leading to poor translation quality. Previous research has shown that document-level NMT, comprising of sentence and context encoders and a decoder, leverages context from neighbouring sentences and helps improve translation quality. In simultaneous translation settings, the context from previous sentences should be even more critical. To this end, in this paper, we propose `wait-k` simultaneous document-level NMT where we keep the context encoder as it is and replace the source sentence encoder and target language decoder with their `wait-k` equivalents. We experiment with low and high resource settings using the Asian Language Treebank (ALT) and OpenSubtitles2018 corpora, where we observe minor improvements in translation quality. We then perform an analysis of the translations obtained using our models by focusing on sentences that should benefit from the context where we found out that the model does, in fact, benefit from context but is unable to effectively leverage it, especially in a low-resource setting. This shows that there is a need for further innovation in the way useful context is identified and leveraged.

1 Introduction

Neural machine translation (NMT) (Bahdanau et al., 2015; Luong et al., 2016) is an end-to-end approach known to give the state of the art results for a variety of language pairs. In standard NMT, the entire source language sentence is fed to the model, and once the entire target language sentence is generated, it is presented to the user. However, in a real-time translation setting, translation models are expected to present translated words or phrases as they are generated. Furthermore, waiting for the entire source language sentence adds to the latency, and therefore an optimal solution is to have a model that can start generating target language words right after the first few source language words are available for translation. This is known as simultaneous NMT (SNMT) and is known for its poor translation quality, especially in low-resource settings. The concept of waiting for `k` words or tokens before generating target language words or tokens is known as `wait-k` SNMT (Ma et al., 2019). In this paper, we work with the Transformer architecture as the standard NMT model, consisting of a bidirectional encoder and unidirectional decoder. The decoder is able to attend to all source language tokens when generating target language tokens. However, in the case of the `wait-k` SNMT model, the standard encoder and decoder are replaced with their SNMT equivalents, which are a unidirectional encoder and a modified decoder, respectively. The decoder can only look at $i + k - 1$ encoder tokens when predicting the i^{th} token. We are aware of a previous work that has shown that using an image as an additional modality can help improve translation quality in a `wait-k` setting when `k` is a small value around 1 to 4 (Imankulova et al., 2020; Caglayan et al., 2020). The additional image modality provides the model with a form of *context* which helps disambiguate hard-to-translate phenomena, especially when needed information is not available yet during translation. An additional image modality may not always be available, and thus, taking advantage of the context in the form of previously seen sentences is the only viable option.

Research in document-level NMT has already proven that context from neighbouring sentences can help enhance representations and thereby improve translation quality (Tiedemann and Scherrer, 2017; Jean et al., 2017; Wang et al., 2017). The simplest document-level NMT architecture involves using an additional encoder that encodes the context sentences, following which the encoded context is used to augment the representation of the sentence to be translated (Zhang et al., 2018). Just like using an image as a modality helps enrich the encoding of the sentence with additional disambiguation information, the context sentences might also contain such useful information. We already know that in an SNMT setting, due to partial sentences being translated, the amount of context available to the decoder is limited, and thus leveraging the context sentences should significantly boost SNMT translation quality. This motivated us to combine document-level NMT with SNMT leading to document-level SNMT.

Our document-level SNMT architecture is simple, where we have a sentence encoder, context encoder, and a decoder except that the sentence encoder and decoder are `wait-k` SNMT equivalents of the standard encoder and decoder. We experiment with a high-resource OpenSubtitles2018 dataset for English→Russian and Russian→English translation and a low-resource ALT document-level dataset for English→Japanese and Japanese→English translation. Our observations show that document-level context helps improve translation slightly in both settings but not by a large margin. We then perform a statistical and manual analysis of the translations where we observe that while SNMT models definitely benefit from context, they are unable to utilize context effectively and sometimes suffer due to the provided context. This opens up the possibility of research into better mechanisms for leveraging context more effectively.

2 Related Work

For simultaneous translation, it is crucial to predict the words that have not appeared yet. Mainly, SNMT can mostly be implemented with fixed or adaptive policies (Zheng et al., 2019b). Adaptive policy decides whether to READ another source word or WRITE a target word in one model (Grissom II et al., 2014; Matsubara et al., 2000; Oda et al., 2015). Most dynamic models with adaptive policies (Gu et al., 2017; Dalvi et al., 2018; Zheng et al., 2019a,c, 2020a) focus on mechanisms that determine the optimal number of source language tokens to wait for before generating the next target language token. Meanwhile, Ma et al. (2019) proposed a simple `wait-k` method with fixed policy, where the decoder starts generating the target language tokens the moment `k` source language tokens are available. However, their model for simultaneous translation relies only on the source sentence. This research concentrates on the `wait-k` approach leveraging document-level information from previous context sentences.

Document-level NMT leverages context beyond the current sentence in order to improve translation quality (Tiedemann and Scherrer, 2017; Jean et al., 2017; Wang et al., 2017; Voita et al., 2018, 2019; Zheng et al., 2020b; Fernandes et al., 2021). Document-level NMT models can be implemented as a post-processing model or context-aware model. The post-processing models use an additional module to use context on generated translations (Xiong et al., 2019; Voita et al., 2019). However, post-processing generated translations may lead to higher latency, which is counter-intuitive in a simultaneous translation scenario. On the other hand, context-aware models leverage additional context during translation. For example, Tiedemann and Scherrer (2017) proposed to simply concatenate the previous sentences in both the source and target side to the input to the system. Jean et al. (2017); Bawden et al. (2018); Zhang et al. (2018) use separate context encoder for a few previous source sentences. Similarly, we also use a separate context encoder to extract document-level information. However, we incorporate document-level information into SNMT in order to improve translation quality, where only information from the source sentence is insufficient during translation.

3 Methods

3.1 Background: Wait-k Simultaneous NMT

The most straightforward approach for SNMT is the `wait-k` approach (Ma et al., 2019) with a fixed policy. As tokens are fed to the encoder one at a time, we have to rely on a unidirectional encoder that cannot attend to future tokens. Once the encoder has been fed `k` tokens, the decoder starts generating a token at a time. This means that at the i^{th} decoding step, the encoder and decoder can only see the first $\mathtt{k} + i - 1$ encoder token representations. Once the whole input sentence is available, `wait-k` behaves like regular NMT except with a unidirectional encoder. Different from (Ma et al., 2019) we have a unidirectional encoder, so when a new source token arrives, the encoder representations for the previous tokens are not updated. This can have a minor impact on the overall translation quality, but this paper aims to understand how context affects SNMT.

3.2 Background: Document-level NMT

Suppose X, X_c and Y are the source sentence, context sentences, and the target sentence. In this paper, we work with SNMT, and hence X_c only consists of past sentences, which for simplicity we concatenate into a single long context sentence[1]. Document-level NMT involves using X and X_c together for translation. In the case when only X and Y are available, X is fed to an encoder (E), leading to a sentence encoding $E(X)$. This sentence encoding is then attended to by the decoder in order to produce the translation $Y^{'} = D(E(X))$. When X_c is available we encode it using a context encoder (E_c) leading to context encoding $E_c(X_c)$ which is then used for translation along with $E(X)$ as $Y^{'} = D(E(X), E_c(X_c))$. It is a common practice to share the parameters of the sentence and context encoders. A key component of document-level NMT is the incorporation of $E_c(X_c)$ into the framework by combining it with $E(X)$. This paper considers two simple approaches, which we dub as “multi-source” (MS) and “context-attention” (CA).

3.2.1 MS: Multi-Source Based Context Incorporation

This method treats the context as an additional source of information similar to the setting in multi-source NMT (Zoph and Knight, 2016; Dabre et al., 2017). In multi-source NMT, the decoder is modified to attend to multiple source sentences, and this approach should help incorporate context into the decoding process. For vanilla NMT, the cross attention mechanism of the decoder takes in $E(X)$ and produces a weighted representation, the attention, A. Given the context encoding $E_c(X_c)$ we additionally compute the context attention A_c. We combine A and A_c into A_{comb}, the context augmented attention, using a simple gating mechanism as $A_{comb} = \alpha * A + (1 - \alpha) * A_c$ where $\alpha = sigmoid(W_{comb} * [A : A_c])$. [:] indicates concatenation of representations along the hidden layer axis. W_{comb} is the weight matrix of size

[1] This means that the memory requirements will increase, but we believe that this is an acceptable trade-off if translation quality improves. Furthermore, we can use sequence distillation Kim and Rush (2016) to compress these models, which have a smaller memory footprint

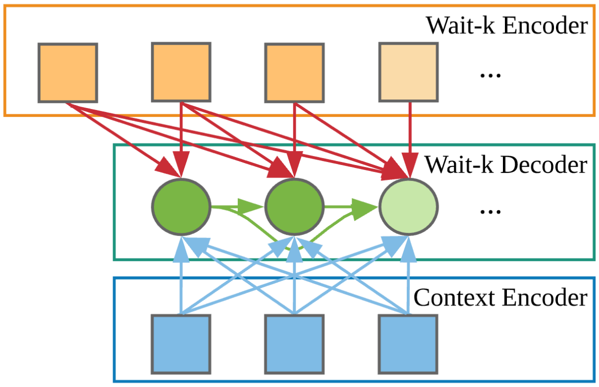


Figure 1: A simplified overview of our simultaneous document level NMT model which uses previous source sentences as context.

$[2h, h]$ where h is the model's hidden size. α is a weight that can help interpolate A and A_c to determine the balance between them.

3.2.2 CA: Context Attention Based Context Incorporation

This method is same as the one in Voita et al. (2018). Where the multi-source approach involves combining $E(X)$ and $E_c(X_c)$ in the decoder by combining the attentions obtained from them (A and A_c), this approach combines $E(X)$ and $E_c(X_c)$ into a single $E_{comb}(X, X_c)$ which is then fed to the decoder. Thus, the decoder sees one encoder representation instead of two.

To combine $E(X)$ and $E_c(X_c)$, $E(X)$ is fed to a self-attention layer which gives $E_{sa}(X)$ and $E_c(X_c)$ is fed to a cross-attention layer where EX is the query and $E_c(X_c)$ is the key/value which gives $E_{ca,c}(X_c)$. By doing so, $E_{sa}(X)$ and $E_{ca,c}(X_c)$ have the same shape and can be combined via the gating mechanism in the previous section into $E_{comb}(X, X_c)$.

Apart from these two combination methods, there are several others (Libovický et al., 2018) which we will explore in the future.

3.3 Our Method: Document-level SNMT

Document-level NMT can be easily extended to document-level SNMT by enforcing the SNMT constraint on the sentence encoder E and the sentence cross-attention mechanism A. No such constraints are placed on the context encoder E_c. Refer to Figure 1 for a simple overview of our method. It shows that at the i^{th} decoding step, the decoder and encoder can access the context representations fully but only $k + i - 1$ source sentence representations.

4 Experimental Settings

We describe experimental settings aimed at helping verify the degree to which document context helps improve translation quality in a simultaneous translation setting.

4.1 Datasets and preprocessing

We experimented with English→Russian and Russian→English translation using a corpus created by (Voita et al., 2018), derived from the OpenSubtitles2018 corpus, consisting of 1.5M training sentences where each sentence has 3 sentences as context. The development and test sets consist of 10,000, 4 sentence documents leading to a total of 40,000 sentences which can have up to 3 context sentences. This dataset belongs to the spoken language domain, where we expect that document context should be very helpful in improving translation quality. Given that Russian has flexible word order, missing information in an incomplete source sentence can be complemented via the context. We also experimented with the low-resource Asian Language Treebank (ALT)dataset (Riza et al., 2016), which contains sentence level aligned document pairs split into training/development/test sets of 18,088/1,000/1,018 lines spanning 1,698/98/97 documents, respectively. We experimented with English→Japanese and Japanese→English translation. Japanese has subject-object-verb word order, whereas English has subject-verb-object, so we expect document context to be helpful whenever the object or verb-related information is missing for incomplete sentences in an SNMT setting.

Regarding preprocessing, we segmented the Japanese source sentences using MeCab, and our NMT implementation handles other preprocessing, such as subword tokenization. When providing document context sentences to our models, we concatenate previous N context sentences to form a single long sentence before feeding it to the model along with the sentence to be translated. Naturally, the first sentence of the document will have no context sentence, which we designate with a special token $< EMPTY >$.

4.2 Implementation and Training Details

We modified the Transformer (Vaswani et al., 2017) implementation in tensor2tensor v1.15.4[2], which has an internal subword segmentation mechanism. We set the separate source and target subword vocabulary sizes of 8,000 for the ALT dataset and 32,000 for the OpenSubtitles2018 dataset. We use hyperparameters of the "transformer_base" model for English→Russian and Russian→English translation whereas for English→Japanese and Japanese→English translation we use the "transformer_base_single_gpu" model hyperparameters. The "transformer_base" models are trained on 8 NVIDIA V100 GPUs, whereas the"transformer_base_single_gpu" models are trained on a single NVIDIA V100 GPU. We save and evaluate our models on the development set every 1000 batches with BLEU (Papineni et al., 2002) as the evaluation metric. We train our models till the BLEU score does not increase for ten consecutive evaluations. We average the last ten saved checkpoints and then decode the model. As we work in a simultaneous translation setting, greedy search makes sense as tokens should be output one at a time [3].

4.3 Models Compared

We train and compare the following types of full sentence and `wait-k` SNMT models for both datasets:
1. Non-contextual models: where the document context is not used
2. Contextual models: which use up to N previous sentences as context. $N = 1$ for English↔Japanese[4] and $N = 1, 2, 3$ for English↔Russian.

5 Results

We describe the results of our experiments in resource-rich and resource-poor settings.

[2] `https://github.com/tensorflow/tensor2tensor/tree/v1.15.4`

[3] It's possible to consider a sophisticated beam search method, but that is beyond the scope of this paper.

[4] In reality, we had experimented with $N = 2$, but found out that the translation quality, measured in BLEU, dropped. We suspect that this is because either the model ends up paying unnecessary attention to the context or that the low-resource setting hinders the model from learning how to utilize context effectively. Ultimately we feel that $N = 1$ is a practical choice for the ALT dataset because it contains sentences with around 20 words on average. The longer the context sentence, the more computations the cross attention mechanism has to make, which slows decoding, which is ultimately what we are trying to avoid via SNMT while incorporating context. We were able to consider all 3 context sentences for English↔Russian because each sentence was substantially smaller, which does not impact decoding time as badly. In the future, we can consider sparse attention mechanisms such as locality sensitive hashing, which is used in the Reformer (Kitaev et al., 2020).

			Russian→English				English→Russian			
Model	wait-k	CT	CS=0	CS=1	CS=2	CS=3	CS=0	CS=1	CS=2	CS=3
Full Sentence	-	MS	34.9	35.2	35.5	**35.7**	26.7	27.0	**27.2**	**27.2**
	-	CA	34.9	35.3	**35.8**	35.6	26.7	27.0	27.2	**27.5**
SNMT	1	MS	23.5	23.6	24.0	**24.1**	13.2	13.4	13.4	**13.5**
	1	CA	23.5	23.7	23.8	**24.1**	13.2	13.3	**13.4**	13.3
	2	MS	28.8	28.9	**29.4**	29.3	17.6	17.7	**18.0**	17.9
	2	CA	28.8	29.1	**29.5**	**29.5**	17.6	17.9	18.0	**18.1**
	4	MS	32.9	33.2	33.5	**33.7**	23.7	23.7	23.7	**23.9**
	4	CA	32.9	33.1	**33.6**	**33.6**	23.7	23.6	**23.8**	**23.8**
	6	MS	33.9	34.3	34.5	**34.8**	25.7	25.7	25.8	**26.0**
	6	CA	33.9	34.4	34.6	**34.8**	25.7	25.7	25.9	**26.3**
	8	MS	34.3	34.6	35.0	**35.3**	26.2	26.3	26.5	**26.8**
	8	CA	34.3	34.8	34.9	**35.1**	26.2	26.4	26.5	**26.8**

Table 1: BLEU scores for English→Russian and Russian→English translation using the OpenSubtitles2018 corpus. Results are presented for full sentence and SNMT models using either no context or up to 3 context sentences ($CS = 0, 1, 2, 3$). CT indicates the document context incorporation technique which can be MS (Multi-Source) or CA (Context Attention). As improvements greater than 0.1 BLEU are statistically significant and most cases show improvement over baselines, we do not mark all significantly improved scores to avoid cluttering. For each type of model (full sentence or wait-k) for a language pair, we mark the best scores in bold.

			Japanese→English		English→Japanese	
Model	wait-k	CT	CS=0	CS=1	CS=0	CS=1
Full Sentence	-	MS	8.8	**9.0**	13.7	**14.1**
	-	CA	**8.8**	8.6	13.7	**14.2**
SNMT	1	MS	3.1	**3.2**	**9.3**	9.1
	1	CA	3.1	**3.3**	**9.3**	8.7
	2	MS	**3.8**	3.7	**10.4**	9.6
	2	CA	**3.8**	3.7	**10.4**	10.0
	4	MS	**4.8**	4.7	**12.1**	11.7
	4	CA	**4.8**	4.7	**12.1**	11.3
	6	MS	5.5	**5.6**	12.9	**13.0**
	6	CA	5.5	**5.6**	**12.9**	**12.9**
	8	MS	5.9	**6.3**	13.6	**13.7**
	8	CA	5.9	**6.5**	**13.6**	13.2

Table 2: BLEU scores for English→Japanese and Japanese→English translation using the ALT corpus. Results are presented for full sentence and SNMT models using either no context or up to 1 context sentence ($CS = 0, 1$). CT indicates the document context incorporation technique which can be MS (Multi-Source) or CA (Context Attention). For each type of model (full sentence or wait-k) for a language pair, we mark the best scores in bold.

5.1 Resource Rich English↔Russian translation

Table 1 gives the BLEU scores for English↔Russian translation.

5.1.1 Non-contextual: Full Sentence versus SNMT models

Regarding the baselines, it is clear that the SNMT models with small `wait-k`'s give poor translation quality as compared to the full sentence models. Increasing the value of `wait-k` naturally improves the translation quality, where a value of `k` = 8 leads to results that are within 1 BLEU of the results of the full sentence models. Given that the average sentence length for the Russian–English dataset is approximately 8 words, it makes sense that $K = 8$ would give the best results.

5.1.2 Context incorporation technique: Multi-Source (MS) versus Context Attention (CA)

The results show that there is no clear answer as to which of MS or CA is superior, which makes both viable solutions for incorporating context into the NMT model. For the remainder of the results section, the BLEU scores we quote will be for the MS approach. Looking at the results, it will be clear that the trends in the improvement of translation quality by incorporating context are similar regardless of the use of MS or CA.

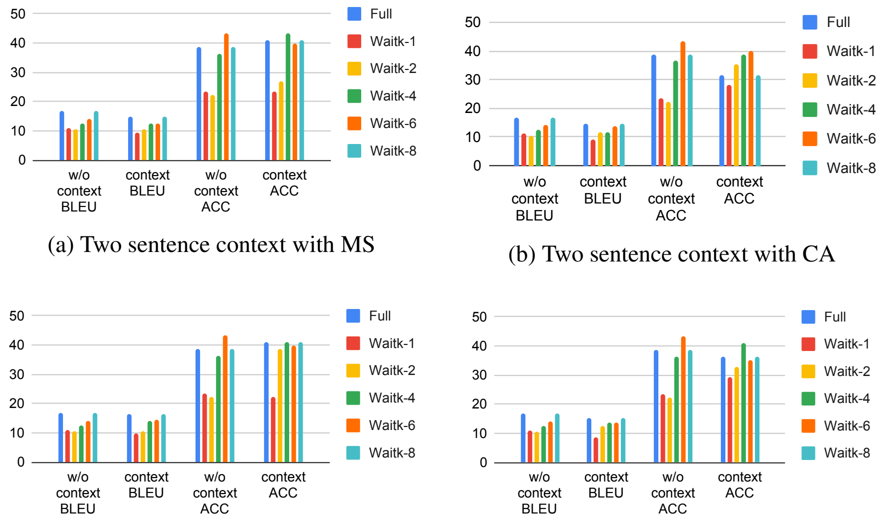
5.1.3 Non-contextual versus Contextual Full-Sentence models

Next, when context sentences are used for full sentence translation for Russian→English, the quality for when up to 1, 2, and 3 previous sentences as context are used is 35.2, 35.5, and 35.7, respectively. Compared to a baseline score of 34.9, the improvements are 0.3, 0.6, and 0.8 BLEU. Similarly, for English→Russian, compared to a baseline score of 26.7, using up to 1, 2, and 3 previous sentences as context lead to translation quality improvements of 0.3, 0.5, and 0.5, respectively. We performed statistical significance testing (Koehn, 2004) which showed that all improvements are significant[5] at $p < 0.05$. This shows that context certainly helps in a spoken language domain, and as the number of context sentences grows, the translation quality also grows steadily.

5.1.4 Non-contextual versus Contextual SNMT models

Comparing the `wait-k` non-contextual model against contextual models using up to N context sentences shows that, once again, context is helpful in an SNMT setting. When using up to 3 context sentences, for `wait-k` values of 1, 2, 4, 6 and 8, the BLEU score improvements over their non-contextual counterparts are 0.6, 0.5, 0.8, 0.9, 1.0, respectively, for Russian→English translation. Similarly for the reverse direction the improvements are 0.3, 0.3, 0.2, 0.3, 0.6. One important observation is that the improvements are almost proportional to the value of `wait-k`. As we wait for more source language tokens, the impact of the previous sentences as context seems to be higher. This makes sense because the importance of the context is determined using a gating mechanism, and the more information we have about the current sentence, the better the gating mechanism will be at determining what part of the context should be used. Finally note the maximum gain for SNMT models using up to 3 context sentences which is 1.0 for Russian→English and 0.6 for English→Russian. Compared to the full sentence models, the corresponding gains are 0.8 and 0.5. Previously we have seen that a difference of 0.1 BLEU is sufficient for it to be statistically significant, which means that SNMT models experience significantly larger improvements in translation quality when compared to their full sentence counterparts.

[5] Note that the test set contains 40,000 sentences, so even a small improvement of 0.1 BLEU will be significant.



(a) Two sentence context with MS

(b) Two sentence context with CA

(c) One sentence context with MS

(d) One sentence context with CA

Figure 2: BLEU and accuracy (ACC) results for models using one or two previous sentences as context. We perform analyses for the multi-source (MS) and context-attention (CA) based context incorporation mechanisms.

5.2 Resource Poor English↔Japanese translation

Table 2 gives the BLEU scores for English↔Japanese translation. Looking at the absolute BLEU scores shows that context does lead to minor improvements in translation quality regardless of a full sentence or SNMT models. Unfortunately, the improvements are not statistically significant. Although we do not show it here, using additional context sentences led to a drop in translation quality. We suppose that this may be either due to the low-resource nature of the ALT dataset or perhaps there are not many cases where context should be helpful. Note that our context NMT model takes a weighted average of the attentions of the current and the context sentence, and so the translation quality may degrade if there are very few cases where context is needed. To this end, we decided to perform a statistical and manual analysis of the models for English→Japanese translation.

6 Analysis

6.1 Translation of Context-Aware Tokens

We investigate whether SNMT performance is improved by using contextual information. Therefore, we created context-aware parallel data in which the target sentence contains the tokens related to the previous target sentence. For example, given the context source sentence “The 2008 Taipei Game Show, organized by the Taipei Computer Association (TCA), ended on Monday, and was different from shows of past years.”, and the source sentence “This could be seen in the gaming population, industry, and exhibition arrangements.” in a simultaneous manner, the generated target sentence should be “ゲーム の 人口 、 産業 、 そして 展示 会 の 配列 で 見る こと が でき た 。”. Here, “ゲーム” means “game” and it is a token related to the context. The context sentence contains information about the game, and this can help translate “ゲーム” that appears at the beginning of the target sentence, where it is not available yet from the source sentence (e.g., $\texttt{k} < 6$). We randomly investigated such sentence pairs from the

English→Japanese	
Context	Mr. Bush's talks with **Saudi** leaders also are expected to cover arms sales.
Source	Before heading to **Saudi Arabia**, Mr. Bush visited Dubai briefly.
Target	**サウジアラビア** に 向かう 前 に 、 ブッシュ 氏 は ドバイ を 短 期間 訪問 し た 。
`wait`-2 w/o context	既に 割れ て いる 前 に 、 ブッシュ 氏 は ドバイ に つい て 言及 し た 。 (Before it was already cracked, Mr. Bush mentioned Dubai.)
`wait`-8 w/o context	**サウジアラビア** の 王室 に 証言 する 前 に 、 ブッシュ 氏 は ドバイ へ の 説明 を 訪問 し た 。 (Before testifying before the **Saudi** royal family, Mr. Bush visited Dubai to explain.)
Full w/o context	**サウジアラビア** の 王室 に 証言 する 前 に 、 ブッシュ 氏 は ドバイ へ の 説明 を 訪問 し た 。 (Before testifying before the **Saudi** royal family, Mr. Bush visited Dubai to explain.)
`wait`-2 w/ context	**サウジアラビア** の イスラム 教 徒 の 前 に 、 ブッシュ 氏 は 先週 記者 を 訪問 し た 。 (Before the **Saudi** Muslims, Mr. Bush visited the press last week.)
`wait`-8 w/ context	**サウジアラビア** の 王室 に 向 かっ て 前 に 、 ブッシュ 氏 は ドバイ を 訪問 し た 。 (Before heading to the royal family in **Saudi Arabia**, Mr. Bush visited Dubai.)
Full w/ context	**サウジアラビア** の 王室 に 向 かっ て 前 に 、 ブッシュ 氏 は ドバイ を 訪問 し た 。 (Before heading to the royal family in **Saudi Arabia**, Mr. Bush visited Dubai.)

Table 3: Translation examples generated by non-contextual models as well as the contextual models using one previous sentence as context and the multi-source (MS) context incorporation method. Sentences in parentheses are the English meanings of the translation results.

test data of WAT data and extracted 50 of them. Using BLEU and accuracy, calculated by the sum of correctly translated sentences that include the token that needs context to be translated, divided by the number of sentences, we evaluate whether the performance of the SNMT model is improved by using the context.

Figure 2 shows that BLEU and accuracy results for contextual models, using up to one or two previous sentences[6] as context, for created context-aware parallel data. In BLEU, it can be seen that the results are almost the same between the non-contextual and the contextual models. On the other hand, the results of accuracy differ between the non-contextual and the contextual models. In particular, accuracy is improved by considering the context at k = 1, 2, and 4. From this result, it can be seen that tokens related to the context can be translated by considering the context in SNMT. Our analysis also leads us to believe that it is difficult for BLEU to evaluate the improvement due to the context because BLEU was not designed in that way. This shows

[6] We have mentioned earlier that using two sentences as context led to a drop in translation quality but our analysis shows that they help provide context that is useful despite lowering the overall translation quality.

that there is a need for context-aware evaluation mechanisms.

6.2 Examples of Translations

In order to understand how the translation quality is improved by using context, we analyze the following translations: Table 3 shows the translation examples generated by non-contextual as well as the contextual models using one previous sentence as context and the multi-source (MS) context incorporation method. The "Saudis" contained in the context sentence is thought to be helpful when translating "サウジアラビア" which means "Saudi Arabia" in the source sentence. If `k` is 4 or less, "Saudi Arabia" will not be seen by the decoder. Since the translation result of `k` = 8 and the full sentence is the same, it can be seen that the effect of the missing words is almost eliminated when `k` is large in `wait-k`. "Saudi Arabia" was not translated with `k` = 2 without context, but it was correctly translated using the contextual model. From this translation example, we can see that the context helps to translate the words related to it. However, given that the overall corpus level BLEU does not show a large amount of improvement, we suspect that the current context incorporation mechanisms are not good at determining when the context should and should not be used. This means that we need to design better context relevance mechanisms.

7 Conclusion

We proposed `wait-k` document-level simultaneous NMT to complement the information of incomplete input during the translation process. Our proposed method is to replace the source encoder and target language decoder with `wait-k` equivalents while keeping the context encoder. The experimental results show that the proposed method slightly improves the translation quality in high-resource settings but not by appreciable amounts in low-resource settings. The analysis showed that `wait-k` models are more context-aware and rely on context whenever it should be helpful. However, the current model is unable to successfully determine when the context should be used, preventing the successful utilization of context. This indicates that we need to investigate further more effective ways to utilize the previous sentences in the document as context. Our human evaluation was also rather limited, and in the future, we plan to conduct a human evaluation to determine which kind of context-aware phenomena (pronoun disambiguation, word sense disambiguation) our approaches can address.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Caglayan, O., Ive, J., Haralampieva, V., Madhyastha, P., Barrault, L., and Specia, L. (2020). Simultaneous machine translation with visual context. In *EMNLP*, pages 2350–2361. Association for Computational Linguistics.

Dabre, R., Cromieres, F., and Kurohashi, S. (2017). Enabling multi-source neural machine translation by concatenating source sentences in multiple languages. In *Proceedings of MT Summit XVI, vol.1: Research Track*, pages 96–106, Nagoya, Japan.

Dalvi, F., Durrani, N., Sajjad, H., and Vogel, S. (2018). Incremental decoding and training methods for simultaneous translation in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 493–499. Association for Computational Linguistics.

Fernandes, P., Yin, K., Neubig, G., and Martins, A. F. (2021). Measuring and increasing context usage in context-aware machine translation. *arXiv preprint arXiv:2105.03482*.

Grissom II, A., He, H., Boyd-Graber, J., Morgan, J., and Daumé III, H. (2014). Don't until the final verb wait: Reinforcement learning for simultaneous machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1342–1352.

Gu, J., Neubig, G., Cho, K., and Li, V. O. (2017). Learning to translate in real-time with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1, Long Papers)*, pages 1053–1062.

Imankulova, A., Kaneko, M., Hirasawa, T., and Komachi, M. (2020). Towards multimodal simultaneous neural machine translation. In *WMT*, pages 594–603. Association for Computational Linguistics.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017). Does neural machine translation benefit from larger context? *arXiv preprint arXiv:1704.05135*.

Kim, Y. and Rush, A. M. (2016). Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.

Kitaev, N., Kaiser, Ł., and Levskaya, A. (2020). Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Libovický, J., Helcl, J., and Mareček, D. (2018). Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. Association for Computational Linguistics.

Luong, M., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2016). Multi-task sequence to sequence learning. In *Proceedings of International Conference on Learning Representations*.

Ma, M., Huang, L., Xiong, H., Zheng, R., Liu, K., Zheng, B., Zhang, C., He, Z., Liu, H., Li, X., Wu, H., and Wang, H. (2019). STACL: Simultaneous translation with implicit anticipation and controllable latency using prefix-to-prefix framework. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3036.

Matsubara, S., Iwashima, K., Kawaguchi, N., Toyama, K., and Inagaki, Y. (2000). Simultaneous Japanese-English interpretation based on early prediction of English verb. In *Proceedings of The Fourth Symposium on Natural Language Processing*, pages 268–273.

Oda, Y., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. (2015). Syntax-based simultaneous translation through prediction of unseen syntactic constituents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 198–207.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Riza, H., Purwoadi, M., Gunarso, Uliniansyah, T., Ti, A. A., Aljunied, S. M., Mai, L. C., Thang, V. T., Thai, N. P., Sun, R., Chea, V., Soe, K. M., Nwet, K. T., Utiyama, M., and Ding, C. (2016). Introduction of the Asian language treebank. In *Proc. of O-COCOSDA*, pages 1–6.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Proceedings of the Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Voita, E., Sennrich, R., and Titov, I. (2019). Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 877–886, Hong Kong, China. Association for Computational Linguistics.

Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017). Exploiting cross-sentence context for neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2826–2831, Copenhagen, Denmark. Association for Computational Linguistics.

Xiong, H., He, Z., Wu, H., and Wang, H. (2019). Modeling coherence for discourse neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7338–7345.

Zhang, J., Luan, H., Sun, M., Zhai, F., Xu, J., Zhang, M., and Liu, Y. (2018). Improving the transformer translation model with document-level context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 533–542, Brussels, Belgium. Association for Computational Linguistics.

Zheng, B., Zheng, R., Ma, M., and Huang, L. (2019a). Simpler and faster learning of adaptive policies for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1349–1354.

Zheng, B., Zheng, R., Ma, M., and Huang, L. (2019b). Simultaneous translation with flexible policy via restricted imitation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5816–5822.

Zheng, R., Ma, M., Zheng, B., and Huang, L. (2019c). Speculative beam search for simultaneous translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1395–1402.

Zheng, R., Ma, M., Zheng, B., Liu, K., and Huang, L. (2020a). Opportunistic decoding with timely correction for simultaneous translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 437–442.

Zheng, Z., Yue, X., Huang, S., Chen, J., and Birch, A. (2020b). Towards making the most of context in neural machine translation. In *IJCAI*.

Zoph, B. and Knight, K. (2016). Multi-source neural translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 30–34, San Diego, California. Association for Computational Linguistics.