Ph.D. DISSERTATION

Context-aware Document-level Neural Machine Translation

문맥 인식기반의 문서 단위 신경망 기계 번역 연구

BY

HWANG YONGKEUN

FEBRUARY 2022

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Ph.D. DISSERTATION

Context-aware Document-level Neural Machine Translation

문맥 인식기반의 문서 단위 신경망 기계 번역 연구

BY

HWANG YONGKEUN

FEBRUARY 2022

DEPARTMENT OF ELECTRICAL AND
COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Context-aware Document-level Neural Machine Translation

문맥 인식기반의 문서 단위 신경망 기계 번역 연구

지도교수 정 교 민
이 논문을 공학박사 학위논문으로 제출함

2022년 2월

서울대학교 대학원

전기정보공학부

황 용 근

황용근의 공학박사 학위 논문을 인준함

2022년 2월

위 원 장: _____
부위원장: _____
위 원: _____
위 원: _____
위 원: _____

Abstract

The neural machine translation (NMT) has attracted great attention in recent years, as it has yielded state-of-the-art translation quality. Despite of their promising results, many current NMT systems are sentence-level; translating each sentence independently. This ignores contexts on text thus producing inadequate and inconsistent translations at the document-level. To overcome the shortcomings, the context-aware NMT (CNMT) has been proposed that takes contextual sentences as input. This dissertation proposes novel methods for improving the CNMT system and an application of CNMT. We first tackle the efficient modeling of multiple contextual sentences on CNMT encoder. For this purpose, we propose a hierarchical context encoder that encodes contextual sentences from token-level to sentence-level. This novel architecture enables the model to achieve state-of-the-art performance on translation quality while taking less computation time on training and translation than existing methods. Secondly, we investigate the training method for CNMT models, where most models rely on negative log-likelihood (NLL) that do not fully exploit contextual dependencies. To overcome the insufficiency, we introduce coreference-based contrastive learning for CNMT that generates contrastive examples from coreference chains between the source and target sentences. The proposed method improves pronoun resolution accuracy of CNMT models, as well as overall translation quality. Finally, we investigate an application of CNMT on dealing with Korean honorifics which depends on contextual information for generating adequate translations. For the English-Korean translation task, we propose to use CNMT models that capture crucial contextual information on the English source document and adopt a context-aware post-editing system for exploiting contexts on Korean target sentences, resulting in more consistent Korean honorific translations.

keywords: neural machine translation, document-level translation, contrastive learning, deep neural network
student number: 2014-21693

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recently deep neural networks have achieved remarkable success in the field of natural language processing (NLP), including text classification, summarization, question answering, dialog systems, and so on. Machine translation (MT) is also a classic sub-field in NLP that automates translation between natural languages with computer software. MT has received great attention from academics since MT shares a similar objective with many other NLP and artificial intelligence (AI) tasks, that is to fully understand and resemble the human text (speech) at the semantic level. In addition, translation itself is a difficult task even for humans, achieving good MT performance is thus challenging and would attract many researchers. On the other hand, MT has huge potential for business values as the demands on multilingual content like video streaming and global collaboration amongst institutions and individuals, are still increasing. This huge growth of demands for translation exceeds the capability of human translation, so the needs for developing high-quality MT systems become even more increasing.

In the past years, the majority of MT systems were implemented through statistical machine translation (SMT) which uses statistical models [3, 4] and hand-crafted features to represent translation between bilingual sentence pairs. In contrast, neural machine translation (NMT) systems that are based on novel deep neural network architectures [5, 6, 7] require little to no feature engineering. Because of its simple ar-

chitecture and ability in capturing long dependency in the sentence, NMT has achieved great success on translation quality, and become more popular among the researchers and public [8, 9].

Despite its success, MT systems including NMT are mainly sentence-level systems that translate each sentence in isolation, regardless of their inter-document dependencies. In reality, however, the text does not consist of an isolated, unrelated sentence, but collocated and structured groups of sentences. These sentences are often combined by complex linguistic elements, referred to as the discourse [10]. Ignoring the relationships among these discourse elements results in translations that may looks good but lack crucial properties of the text. In other words, ignoring the document context in translation hinders conveyance of the intended meaning.

Figure 1.1 illustrates one of the such limitation of sentence-level translations [11]. In German, the grammatical genders of a pronoun and its antecedent should agree[1]. Since the "Statue" is feminine, referring pronoun should also be feminine as shown in human translation ("its" translated as "deren"). However, the MT generated pronoun with wrong grammatical gender, as "seine" is masculine.

Source	In fairness, Miller did not attack the statue itself. (...) But he did attack its meaning.
Human	Um fair zu bleiben, Miller griff nicht die Statue selbst an. (...) Aber er griff deren Bedeutung an.
MT	Fairerweise hat Miller die Statue nicht selbst angegriffen. (...) Aber er griff seine Bedeutung an.

Figure 1.1: Grammatical genders of a pronoun *deren* should agree with its antecedent *Statue* (feminine), however sentence-level translation resulted in inadequate choice of pronoun *seine* (masculine).

Figure 1.2 is another example of inconsistencies between translations in isolation. The source Korean sentence is a piece of a press release and showing its English trans-

lations from Google Translate. Even without seeing the human translation, we can easily find that the translations of the same name of the shelter "아동안전지킴이집" (House of children's safety guardian) do not agree with each other. Although all the translations shown above may be perceived as adequate on sentence-level, it contains a word that is inconsistent with the rest of the text, resulting in decreased fluency. From these examples, we can conclude that despite its success in sentence-level performance, the MT system would struggle in achieving human-level translation when it still relies on sentence-level and isolated translation.

Source	서울경찰청은 네이버(주)・서울시와 협력하여 아동의 긴급보호소 역할을 하고 있는 아동안전지킴이집의 위치정보 서비스를 제공한다고 밝혔다. PC나 스마트폰으로 '네이버' 또는 '스마트서울맵(서울시 운영)'의 검색창에 '아동안전지킴이집'을 검색하면 서울 시내에 있는 아동안전지킴이집 위치를 한 눈에 확인할 수 있다. 아동안전지킴이집은 '08년부터 아동범죄 예방정책 일환으로 시행하여 서울시 내 1,357개소가 지정되어 있고, 각종 범죄로부터 아동을 보호하는 역할을 수행하고 있다.
Google Translate	The Seoul Metropolitan Police Agency announced that it would provide the location information service of the Child Safety Keeper House, which serves as an emergency shelter for children, in cooperation with Naver Corporation and the City of Seoul. You can check the location of the child safety guard house in downtown Seoul at a glance by searching for 'child safety guard house' in the search bar of 'Naver' or 'Smart Seoul Map (operated by Seoul City)' with a PC or smartphone. Child Safety Keepers have been implemented as part of the child crime prevention policy since 2008, and 1,357 locations in Seoul have been designated, and they are playing a role in protecting children from various crimes.

Figure 1.2: The Google Translate (retrieved on 24 Nov. 2021) have failed to maintain coherence in translations of the same name of the shelter, "아동안전지킴이집".

To overcome the limitation of sentence-level systems, researchers have pioneered

the use of contextual information extracted from surrounding texts on the source and target document. However, most of the existing works to include document context in SMT failed to yield significant improvements due to the limitations of SMT [13]. Recently, with the success of neural machine translation, many researchers are focusing on context-aware neural machine translation (CNMT).

Researchers have first proposed to model the inter-dependencies among the sentences in a document for implementing CNMT systems. The context can either be in the source or the target language, the first few studies have only exploited source-side context [14, 15, 16, 17]. In another studies, researchers have investigated to incorporate target-side context [18, 19, 20, 21] as well. All of the above researches have been focused on modeling architectures for representing contexts in NMT encoder or decoder. For example, researchers first just concatenated preceding and/or succeeding sentence(s) with source sentence [19] and eventually augmented NMT model with an additional encoder that encodes context sentences and fed into NMT encoder and/or decoder [14, 17]. These approaches are quite effective in improving translation quality when using single context sentences but do not consider incorporating broader context by taking multiple sentences. To overcome this limitation, hierarchical architectures for encoding sentences from token-level to sentence-level have been proposed [16, 22, 23]. By compressing token-level representation into sentence-level, these hierarchical models can more efficiently deal with multiple contextual sentences. However, these approaches still rely on source to context token-level attention mechanisms, increasing the computational complexity and becoming more prone to the data sparsity problem.

This dissertation proposes two approaches for improving the CNMT system, one is introducing a new model architecture and the other is developing a new training process. In addition, this dissertation discusses an application of CNMT method. For a new model architecture, we introduce a hierarchical context encoder (HCE) to tackle shortcomings of existing hierarchical CNMT models [24]. HCE first abstracts

sentence-level information from preceding sentences, and then hierarchically encodes context-level information. Although the HCE eliminates the source-to-context attention at the token level, it instead thoroughly compresses the token-level representation using an attentive weighted sum network and models inter-context dependency by using self-attentional networks in the upper level of hierarchy. This increases the computation speed of the encoder as well as minimizes insufficiency in capturing source-to-context dependency. In the experiment, the proposed model records state-of-the-art performance measured in BLEU score on English-Korean, English-German, and English-Turkish corpus. In addition, we show that our HCE also achieves the best performance in 2 specially designed test suits derived from the same English-Korean corpus; a) a crowd-sourced test set which is designed to evaluate how well an encoder can exploit contextual information, b) an English-Korean pronoun resolution test suite for assessing how the model can discriminate the correct and wrong translation of ambiguous pronouns.

After success in improving a hierarchical encoder for CNMT, we move on to the training method for CNMT. Most of the existing CNMT methods have proposed model architectures that employ structural or input modifications in the base sentence-level NMT model to incorporate context information. For learning the parameters, however, these works follow conventional negative log-likelihood (NLL) minimization with taking context as an extra input. Formally, given the source $\mathbf{x}$, target $\mathbf{y}$, n contextual sentences $C = [\mathbf{c}_1, \cdots, \mathbf{c}_n]$ in the data $\mathcal{D}$, we want to find the optimum parameters θ^* as follows:

$$\theta^* = \arg\min_{\theta} \sum_{(\mathbf{x},\mathbf{y},C)\in\mathcal{D}} -\log P_\theta(\mathbf{y}|\mathbf{x}, C). \tag{1.1}$$

Since the NLL objective function does not directly make use of contextual information, the CNMT model trained in this way tends to exploit context implicitly, such as regularization[25, 26]. To fill the gap, we propose a novel coreference-based contrastive learning method (CorefCL) for training CNMT [27]. The CorefCL consists of data augmentation and contrastive learning scheme based on coreference between

the source and contextual sentences. By corrupting automatically detected coreference mentions in the contextual sentence, CorefCL can train the model to be sensitive to coreference inconsistency. The empirical result demonstrates the superiority of our proposed approach, which consistently improves the pronoun resolution accuracy of several CNMT models as well as the overall translation quality.

Finally, we further investigate an application of the CNMT method that how the CNMT can improve a language-dependent problem. This topic is also closely related to CNMT evaluation because traditional metrics like BLEU [28] or METEOR [29] do not consider such problems, and thus may fail to assess the models especially on longer translations. So recent works have focused on designing test suits for addressing specific discourse phenomena such as pronoun resolution in English-German translation [1]. Many of these discourse phenomena are language-specific and still remain undiscovered in the field of CNMT, especially for Asian languages.

To this end, we investigate how the CNMT method can promote translation improvements of Korean honorifics [30]. Our intuition is that the model should exploit the information such as the relationship between speakers from the surrounding sentences for managing the use of honorific expressions. In contrast to previous studies in this dissertation which only used source-side context, we added the target-side context by adopting a context-aware post-editing (CAPE) technique to refine a set of inconsistent sentence-level honorific translations. For evaluation, we design a heuristic to create honorific-labeled test data and we found that using CNMT outperforms sentence-level NMT baselines both in overall translation quality and honorific translations.

The remaining part of this dissertation is organized as follows. Chapter 2 provides a background on NMT. In chapter 3, we explain the proposed hierarchical encoder for efficiently modeling contextual sentences. Chapter 4 explains a contrastive learning framework for enhancing context-aware NMT models. Further investigation on a context-aware NMT application in Korean honorific translation is discussed in Chapter 5. Finally, the dissertation is concluded in Chapter 6.

Chapter 2

Background: Neural Machine Translation

In this chapter, we briefly overview the neural machine translation (NMT). First we present a short history of NMT. Then we review the problem formulation and core components of NMT including the commonly used encoder-decoder architectures, training/decoding techniques, and evaluation metrics for NMT.

2.1 A Brief History

Prior to the recent resurgence of the neural network, most machine translation research focused on statistical MT (SMT) which models the probability of possible translations [3, 31]. SMT consist of two major components; the translation model (TM) and the language model (LM). The TM represents the probability of translation between words and/or phrases and the LM is used to generate more fluent target sentence. Since both can be trained using only text data, SMT rendered many complex engineering efforts obsolete which existed in traditional approaches like rule-based MT (RBMT) or example-based MT (EBMT).

On the other hand, although there were a few early attempts to incorporate neural networks into the translation process (e.g. [32]), none of the neural-based MT methods were able to get reasonable results beyond toy examples at the time. Few years later,

the resurrection of neural networks in MT was started with the introduction of neural language models [33]. One of the earliest works [34] integrated a neural LM instead of the n-gram LM commonly used in SMT and achieved significant improvements in translation quality. In addition to the neural LM, researchers also studied several other uses of neural networks in SMT including word/phrase re-ordering of translated sentence [35], and replacing/extending the TM [36].

Then efforts to model MT by neural networks eventually moved to (pure) neural machine translation, with increased availability of computational resources like graphics processing unit (GPU) and development of novel deep neural network architectures. The earliest and the most widely used architecture for NMT is a sequence-to-sequence, a.k.a. encoder-decoder model [5, 6] that consist of two neural networks to model the source and target sentences. We will further review the encoder-decoder model later regarding its importance for understanding rest of this dissertation.

Soon after the introduction of encoder-decoder model, NMT has dominated the field of MT research with advances in model architecture like attentional network [7], training methods like back-translation [37], and additional techniques in pre/post-processing. This trend is apparent in the list of submitted MT systems in the shared task of Conference of Machine Translation (WMT), one of the oldest and the most active conference specialized in MT. In 2015, only one system [38] was based on pure NMT. However, starting from 2017 almost all submitted systems were NMT-based [39, 40] since many of the NMT systems have outperformed state-of-the-art SMT systems and the performance of NMT is still increasing.

Currently, NMT research is still progressing rapidly, reflected as the growing number of publications NMT related papers in the past few years (Figure 2.1)[1]. As the number of papers grows, many new research directions are being discovered, and of course, the context-aware NMT is one of the ongoing directions.

[1]Example search query: `https://scholar.google.com/scholar?q=neural+machine+translation&hl=ko&as_sdt=0%2C5&as_ylo=2021&as_yhi=2021`

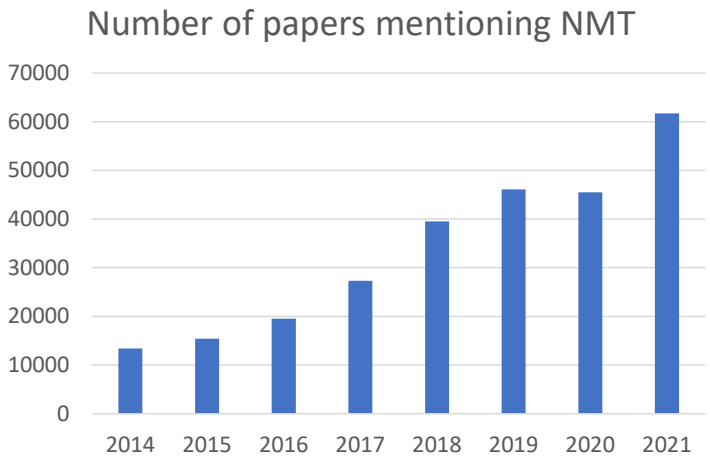

Figure 2.1: Number of papers mentioning "neural machine translation" per year found on Google Scholar (as of 11 Jan. 2022).

2.2 Problem Setup

In this section, we review the problem formulation of NMT. Consider the source sentence $\mathbf{x} = (x_1, \cdots, x_n)$ as an input and the target sentence $\mathbf{y} = (y_1, \cdots, y_m)$ as a output, the goal of sentence-level NMT is to find the most probable target sequence $\hat{\mathbf{y}}$ given a source sentence, that is:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} = P_\theta(\mathbf{y}|\mathbf{x}), \tag{2.1}$$

where θ denotes the trainable set of parameters. The conditional probability $P_\theta(\mathbf{y}|\mathbf{x})$ is modelled using neural networks and can be decomposed as:

$$P_\theta(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{M} P_\theta(y_n|\mathbf{y}_{<n}, \mathbf{x},) \tag{2.2}$$

where y_n is the current target word and $y_{<n}$ are the previously generated words. Note that this decomposition scheme is generally refereed as auto-regressive [41] which models conditional probability of each token (word) as left-to-right causal structure. There are also several implementations of non-auto-regressive NMT (NAT) which

eliminate the causal dependencies and models the whole output tokens at once [42, 43]. However, the following discussion would only focus on auto-regressive methods due to the scope of this dissertation, as most of current CNMT systems are auto-regressive.

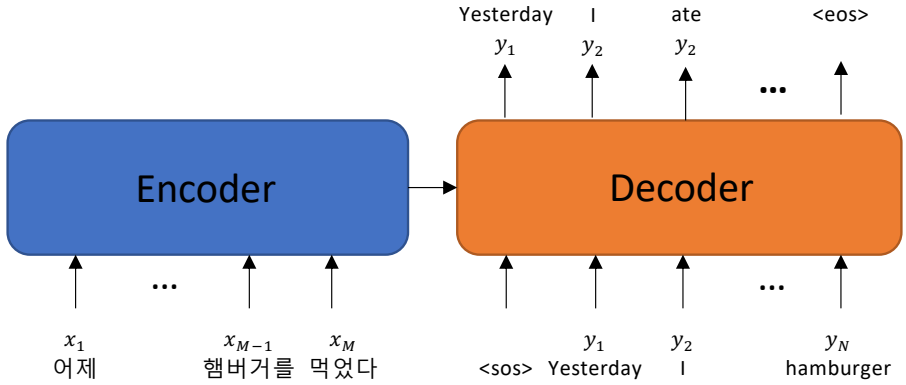

Figure 2.2: An overview of general encoder-decoder model for sentence-level NMT.

2.3 Encoder-Decoder architectures

Generally NMT models are based on an encoder-decoder structure as illustrated in Figure 2.2, where the encoder reads the source sentence to compute a set of vector representation, and the decoder generates the target translation one word at a time given the previously computed source representation. Initially, models used a fixed representation like the last hidden state of encoder network generate the target sentence [5, 6]. It was quickly replaced by the attention-based model [44] that dynamically generates the context representation. These models were based on recurrent neural network (RNN) such as a long short-term memory (LSTM) [45], which is suitable for modelling sequential information. However, RNN's strict sequential computation hindered parallelization within training examples and became a bottleneck when processing long sentences. Recently, a new model architecture based solely on attention mechanisms has been proposed [7]. This self-attentional networks (SANs) removed the recurrence

entirely and has proved to achieve state-of-the-art results on several language-pairs.

There have been also numerous other NMT models such as convolutional neural network (CNN) based models [46] and variations of the SAN like the Universal Transformer [9]. However, we will only cover the RNN and SAN based models in the remaining part of this chapter as these two have been widely used in the field of CNMT. We further review these two models as it is necessary to develop a thorough understanding of rest of this dissertation.

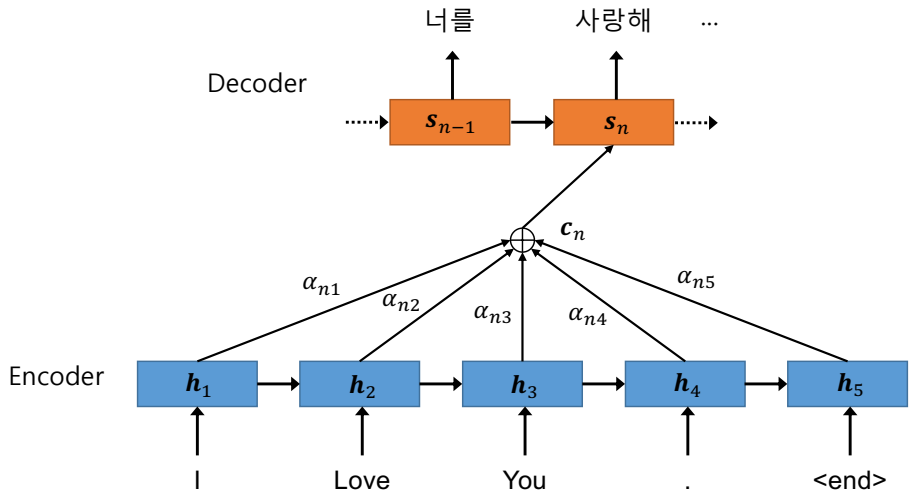

Figure 2.3: RNN-based NMT model with attention mechanism

2.3.1 RNN-based Architecture

In RNN-based NMT models, hidden states of encoder RNN represent individual words of the source sentence. These representation can either be left-to-right direction [6] or bidirectional [44] which consist of the forward and backward RNNs followed by the concatenation of the corresponding bidirectional hidden states. These representations capture information of the corresponding word and its surrounding words in the sentence.

Once the source sentence is processed by encoder RNN, the decoder RNN gen-

erates each words of target sentence given the hidden representations of the source sentence. In the early studies, decoder RNN only have used fixed representation regardless of currently decoding word. With the introduction of attention mechanism, the decoder can dynamically attend to relevant parts of the source sentence at each step of generating the target sentence as shown in Figure 2.3 The context vector or attentional vector $\mathbf{c}_n$ is computed as a weighted summation of the hidden states produced by the encoder RNN, where the weights can be thought of as the alignment probability between a target token (word) at position n and a source symbol at position m. The decoder RNN generates words of the target translation one-by-one in a left-to-right direcetion. The decoder hidden state is computed as follows:

$$\mathbf{s}_n = \mathrm{RNN}(\mathbf{s}_{n-1}, \mathbf{e}_y[y_{n-1}], \mathbf{c}_n), \tag{2.3}$$

where $\mathbf{s}_{n-1}$ is the previous decoder state, $\mathbf{e}_y[y_n]$ is embedding of the word y_n from the embedding table $\mathbf{e}_y$ of the target language, and $\mathbf{c}_n$ is the dynamic context vector that is calculated through attention mechanism:

$$\mathbf{c}_n = \sum_{m=1}^{|\mathbf{x}|} \alpha_{nm} \mathbf{h}_m. \tag{2.4}$$

The weight α_{nm} of each source representation $\mathbf{h}_m$ is computed as:

$$\alpha_{nm} = \frac{e^{f_{nm}}}{\sum_{k=1}^{|\mathbf{x}|} e^{f_{nk}}}, \tag{2.5}$$

where f_{nk} is a scoring function that is generally a feedforward neural network taking $\mathbf{s}_{n-1}$ and $\mathbf{h}_m$ as inputs. This scoring function calculates a alignment score which represent similarity between the query $\mathbf{s}_{n-1}$ and the key $\mathbf{h}_m$. Initially the scoring function is implemented as additive [44] that consist of weighted summation of the query and the key. On the other hand, multiplicative scoring function which calculates similarity using dot-product of the query and the key has also been widely used [47].

The probability of generation of each word y_n is then conditioned on all of the previously generated words $\mathbf{y}_{<n}$ via the state of the RNN decoder $\mathbf{s}_n$, and the source sentence via $\mathbf{c}_n$:

$$\mathbf{u}_n = \tanh(\mathbf{s}_n + \mathbf{W}_{uc}\mathbf{c}_n + \mathbf{W}_{un}\mathbf{e}_y[y_{n-1}]) \tag{2.6}$$

$$P_\theta(\mathbf{y_n}|\mathbf{y}_{<n}, \mathbf{x}) = \text{softmax}(\mathbf{W}_y\mathbf{u}_n + \mathbf{b}_y) \tag{2.7}$$

$$y_n \sim P_\theta(\mathbf{y_n}|\mathbf{y}_{<n}, \mathbf{x}), \tag{2.8}$$

where $\mathbf{W}_{uc}$, $\mathbf{W}_{un}$, $\mathbf{W}_y$ and $\mathbf{b}_y$ are also parameters of the NMT model.

2.3.2 SAN-based Architecture

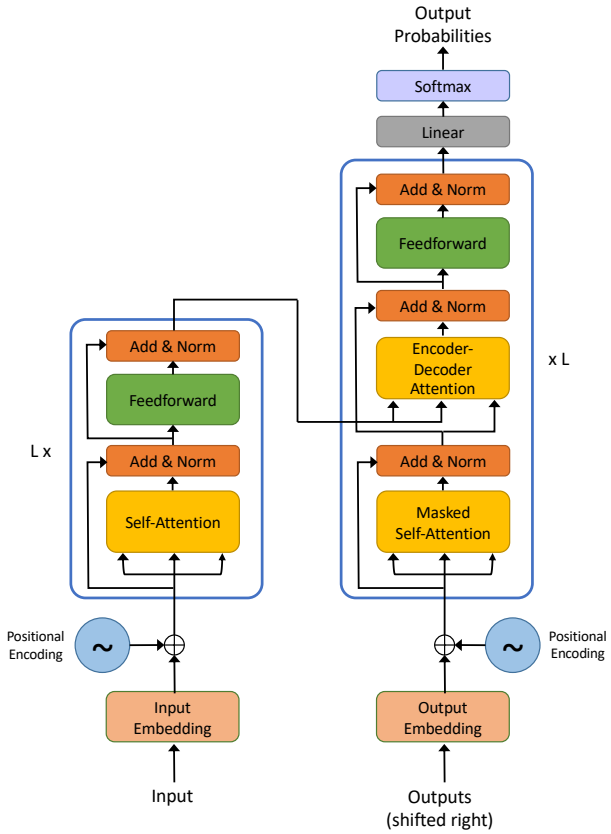

Figure 2.4: Transformer [7], a SAN-based architecture.

RNN-based NMT models have two major limitations. The first limitation is the sequential nature of RNNs. When the model process each input token, the model has to

wait until all previous input tokens have been processed. This can be a bottleneck when processing long sequences. The second limitation is learning long-range dependencies among the tokens . The number of operations required to relate signals from two arbitrary input or output positions grows with the distance between positions, making it difficult to learn complex dependencies between distant positions. The recent self-attentional networks (SAN) like Transformer [7] used stacked self-attention networks followed by point-wise and fully connected layers for both the encoder and decoder for overcoming the limitations of RNN-based NMT models.

The core component of SAN is an attentional network. This can be described as mapping a query ($\mathbf{Q}$) and a set of key-value ($\mathbf{K} - \mathbf{V}$) pairs into an output, where the query, keys, values, and the output are all sequences of vector. Similar to the attention mechanism in RNN-based NMT, it computes a weighted sum of the values as the output, where the weight of each value is computed by a scoring function of the query with the corresponding key.

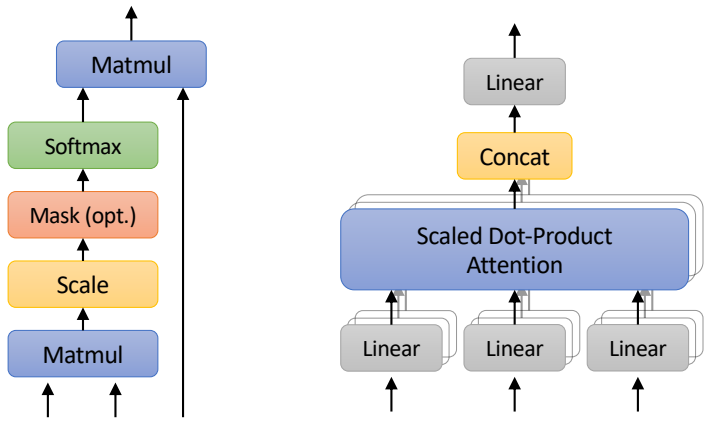

Figure 2.5: Scaled dot-product attention (left) and multi-head attention (right).

The SAN's attentional network consist of the two architectures, scaled dot-product attention and multi-head attention as shown in Figure 2.5. In scaled dot-product atten-

tion, output is computed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d}})\mathbf{V}, \tag{2.9}$$

where d is the hidden diemsion of attentional network and acts as a scaling factor. This scaling factor prevents performance drop especially when d is large. The SAN also performs the multi-head attention, which applies multiple attentional networks to obtain output:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{concatenate}(\text{head}_1, \cdots, \text{head}_h)\mathbf{W}^O \tag{2.10}$$

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_\mathbf{i}^\mathbf{Q}, \mathbf{K}\mathbf{W}_\mathbf{i}^\mathbf{K}, \mathbf{V}\mathbf{W}_\mathbf{i}^\mathbf{V}), \tag{2.11}$$

where $\mathbf{W}_\mathbf{i}^\mathbf{Q}$, $\mathbf{W}_\mathbf{i}^\mathbf{K}$, $\mathbf{W}_\mathbf{i}^\mathbf{V}$, and $\mathbf{W}^O$ are parameter metrices that are also changing dimensions of $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ and the output resepectively. The multi-head attention is beneficial on jointly attending to information from different representation sub-spaces at different positions.

In addition to the attentional network, each of the SAN hidden layers contain point-wise feed-forward network:

$$\text{FNN}(\mathbf{x}) = \text{ReLU}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{2.12}$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$ are all trainable parameters.

In general, the encoder stack of SAN is composed of L identical hidden layers. The attentional network in the encoder stack is a multi-head self-attention allowing each position in the encoder to attend to all positions in the previous layer of the encoder. In this case, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are the same input. On the other hand, the decoder stack has another attentional network, named encoder-decoder attention that is a multi-head attention over the output of the encoder layer. On encoder-decoder attention, the query $\mathbf{Q}$ is an output of self-attentional network and the key-value ($\mathbf{K} - \mathbf{V}$) pairs are obtained from the encoder. In addition, masking is used in the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions.

2.4 Training

To train the model, all parameters in the model are jointly optimized via backpropagation to minimize the loss function over the training set. The loss function is defined as the sum of the negative log-likelihood of predicting a correct symbol y_n in the output sequence for each instance $\mathbf{x}$ in the training set $\mathcal{D}$. Thus, we want to find the optimum set of parameters θ^* as follows:

$$\theta^* = \arg\min_{\theta} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} -\log P_\theta(\mathbf{y}|\mathbf{x}) \tag{2.13}$$

$$= \arg\min_{\theta} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} \sum_{n=1}^{|\mathbf{y}|} -\log P_\theta(y_n|\mathbf{y}_{<n},\mathbf{x}). \tag{2.14}$$

Generally, the model can be trained using random initialization of parameters ("from scratch"). However, when the training data is small, transfer learning from a model that is pre-trained on larger dataset, is adopted to improve the model's performance [48]. Recently, pre-trained language models (PLMs) like BERT [49] and GPT [50] using a variety of unsupervised pre-training methods have shown remarkable success in NLP. Since PLMs are also available in encoder-decoder architectures (e.g. MASS [51], BART [52], and T5 [53]), PLMs are now adopted in several MT transfer learning tasks.

2.5 Decoding

Once the NMT model has been trained, we can use it to translate or decode unseen source sentences. The best output sequence for a given input sequence is produced by:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} = P_\theta(\mathbf{y}|\mathbf{x}), \tag{2.15}$$

Since solving this optimization problem exactly is computationally intractable, approximations such as greedy decoding or beam search are widely used. The basic idea of greedy decoding is to pick the most probable word, i.e. the word having the highest

probability, at each decoding step until the end-of-sentence token is generated. Beam search [54], however, keeps a fixed number of translation hypotheses with the highest log-probability at each time step. A complete hypothesis that is containing the end-of-sentence token is added to the final candidate list. The algorithm then picks the translation with the highest log-probability from this list. If the number of candidates at each timestep is chosen to be one, beam search reduces to greedy decoding. In practice, the translation quality obtained via beam search is significantly better than that obtained via greedy decoding in expense of decoding speed.

2.6 Evaluation

To evaluate the quality of the generated translations, numerous automatic evaluation metrics have been proposed. BLEU [28] has been the most widely used metric for evaluating translation outputs. The main idea of BLEU is to aggregate the count of n-grams that overlap between machine and reference translations. The BLEU metric ranges from 0 to 1, where 1 means an identical output with the reference. Although BLEU correlates well with human judgment, it relies on precision alone and does not take into account recall—the proportion of the matched n-grams out of the total number of n-grams in the reference translation. METEOR [29] was proposed to address the shortcomings of BLEU. It scores a translation output by performing a word-to-word alignment between the translation output and a given reference translation. The alignments are produced via a sequence of word-mapping modules, that is, if the words are exactly the same, same after they are stemmed using the Porter stemmer, and if they are synonyms of each other. After obtaining the final alignment, METEOR computes the parameterized harmonic mean of unigram precision and recall.

Chapter 3

Efficient Hierarchical Architecture for Modeling Contextual Sentences

Recently, interests on context-awareness in neural machine translation tasks have been increasing since additional contextual information is often crucial to produce adequate translations. However, current state-of-the-art translation models including self-attentional networks (SANs) [7] operate on sentence-level do not take account of contextual sentences, hence they record lower performances in spoken languages compared to those in written and formal language documents.

A few studies have addressed this issue by introducing a secondary context encoder to represent contextual sentences then combining them with the source sentence prior to passing them onto the decoder [17, 22, 55]. They proposed context encoders that encode contextual information in the sentence level vectors and use that information in translating input words. These context encoders handle multiple sentences as long word vectors by concatenating them and do not involve the contextual level information.

Such approaches cause critical drawbacks in handling a larger span of contextual sentences. First, the computational complexity of context encoder scales quadratically both with the number of tokens in each contextual sentence and the number of con-

textual sentences. Second, [56, 57] have empirically shown that SAN is limited at capturing long-range dependencies in translation tasks. Hence, concatenating multiple contextual sentences as a long single sentence is not only computationally expensive, but it also weakens the context-awareness of the model for large contexts.

In this chapter, we propose a Hierarchical Context Encoder (HCE) to resolve this issue by hierarchically encoding multiple sentences into a contextual level tensor. HCE first encodes each sentence to a tensor with the SAN encoder, then it converts the encoded tensors into a sentence embedding vector by the attentive weighted summation. Since each sentence embedding vector contains the contextual information of each contextual sentence, we are able to build a context-level tensor by listing all the sentence embedding vectors. Then the context-level tensor is fed into another SAN encoder in order to get a tensor with correlative information between contextual sentences, and the obtained tensor is finally combined with the source encoder to form the final encoder output. Our HCE processes each context sentence separately instead of a long concatenated sentence, hence it shows efficiency in computational complexity. The computational complexity of HCE increases linearly as the number of context sentences increase and HCE shows the fastest running time among standard baseline models in our experiments.

We conduct a series of extensive experiments on NMT with various language pairs to empirically show that our HCE properly yields better translation with multiple context sentences. Our experiments include public OpenSubtitles corpus in English-German, English-Turkish and our web-crawled movie subtitles corpus in English-Korean. On all language pairs, we observed that the translation qualities of our model outperform all the other models measured in BLEU score.

Furthermore, we have constructed an English→Korean evaluation set by crowdsourcing in order to analyze how well our HCE exploits contextual information. Our evaluation set consists of two parts, a part where contextual information is helpful for translation and another part where contextual information is unhelpful. We measure

translation performances in each part and analyze the effects of contextual encoders including HCE by evaluating the performance gap of the two parts. The results from this evaluation set also show that our HCE performs the best among the baseline models. Lastly we create a test suite for pronoun resolution on English→Korean similar to [1, 17]. Evaluation results on the pronoun resolution test suite also reveal the effectiveness of our proposed model.

3.1 Related works

3.1.1 Modeling Context in NMT

Context-aware machine translation models need to focus on additional contexts. In Statistical Machine Translation (SMT), context-awareness is modeled explicitly which is designed for the specific discourse phenomena [13]. For example, anaphora resolution in translation typically involves identifying previously stated nouns, numbers, and genders in source documents and manipulating restoration in target sentences accordingly.

In NMT, either context of the source or the target language can be considered. Exploiting source-side of contexts requires an encoder to represent the multiple context sentence efficiently [17, 22]. On the other hand, the use of target-side contexts often involves multi-pass decoding which translates a part of documents or discourses in the sentence level at first, then refines translations using the previous translations as target contexts [20, 21]. Our proposed model targets to exploit the source side of context-awareness.

The simplest approach to incorporate contexts in the source documents is concatenating all context sentences and passing them into a sentence-level model [18]. In addition, multi-encoder approaches that have an extra encoder for contexts are then introduced. An extra encoder module for context sentences is a natural extension since the source and context sentences do not have the same significance in translation. In

those studies, the context sentences are separately encoded then integrated into the source sentence representations using context-source attention and/or gating network on encoder [17], decoder [14] or both [55].

3.1.2 Hierarchical Context Modeling

The early multi-encoder approaches have inefficiency on modeling broader span of context, since they do not take account of having multiple context sentences. Hierarchical modeling of context sentences is suggested to overcome the inefficiency and capture complex dependencies between a source sentence and context sentences. For example, Wang et. al. [16] uses Recurrent Neural Networks (RNN) encoders operating both on sentence and document level. Miculicich el. al. [22] introduces a hierarchical attention network that encodes context sentences first then summarizes those contexts using a hierarchical structure. Maruf and Haffari [58] introduces a memory network augmented model that summarizes and stores context sentences. Our method is closely related to those approaches, as our proposed encoder also incorporates a hierarchically structured abstraction of encoded context sentences. Maruf et. al. [23] suggests a context attention module which attends to contexts in both word and sentence level. It uses an averaged word embedding as a sentence-level representation, whereas ours generate sentence-level tensor with SAN encoders resulting in richer sentence representation.

3.1.3 Evaluation of Context-aware NMT

On the other hand, how the quality of translation can be benefited with contextual information is a viable research question [14, 19]. Those researches mainly focus on the design of evaluation tasks that assess the performance of the translation model on handling discourse phenomena problems such as pronoun resolution [1, 17]. Voita et. al. [21] also suggests that a carefully designed test suite to evaluate context-aware translation models is crucial since the standard metrics such as BLEU are insensitive on measuring consistency in translation with contexts.

3.2 Model description

In this section, we briefly review common parts of encoders in the context-aware NMT framework. We also review structures of the context-aware encoders which are our baseline models. Then we introduce a detailed structure of our Hierarchical Context Encoder (HCE). In addition, we analyze computational complexities in our proposed encoder and other baseline models.

3.2.1 Context-aware NMT encoders

NMT models without contexts take an input sentence $\mathbf{x}$ in a source language and return an output sentence $\mathbf{y}^*$ in a target language. We denote a target sentence as $\mathbf{y}$ which is used as a golden truth sentence in supervised learning. Each of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{y}^*$ is a tensor that is composed of word vectors, also learnable weights during training.

We especially focus on SAN-based models like Transformer [7] which has recently been widely used in NMT because of its performance and efficiency. Transformer consists of an encoder module and a decoder module, an encoder extract features in $\mathbf{x}$ using self-attention and a decoder generate an output $\mathbf{y}^*$ from the extracted features using both self-attention with itself and attention with the encoder.

Through a single layer in Transformer encoder, an input tensor passes a self-attention layer using multi-head dot product attention and a position-wise feed-forward layer [7]:

$$\mathrm{TransformerEncoder}(\mathbf{x}) = \mathrm{FFN}(\mathrm{MultiHead}(\mathbf{x}, \mathbf{x}, \mathbf{x})). \tag{3.1}$$

The position-wise feed forward layer, denoted as $\mathrm{FFN}(\mathbf{x})$, is composed double linear transformation layer with a ReLU activation as described in Eq. (2.12). The $\mathrm{MultiHead}(\cdot)$ denotes a multi-head dot product attention in Eq. (2.10).

Both the self-attention layer and position-wise feed-forward layer are followed by skip connection and layer normalization. In addition, a stack with multiple $\mathrm{TransformerEncoder}$ is generally used in order to capture more abundant representa-

tions.

With N many additional context sentences $C = [\mathbf{c}_0, \cdots, \mathbf{c}_{N-1}]$ are given, an encoder has to capture contextual information among them then combine the contextual information with source sentence representations. We list four previously suggested models as follows, which are also our baseline models in our experiments;

- **Transformer without contexts (TwoC)**: As a baseline, we have experimented with Transformer without contexts (TwoC) model which has the same structure as [7]. TwoC completely ignores given additional context sentences and only incorporates with the input **x** and the target **y**. The computational complexity is $\mathcal{O}((L_s)^2)$, where L_s is a length of input **x**.

- **Transformer with contexts (TwC)**: The simplest approach is concatenating all context sentences and an input sentence and consider the concatenated sentence as a single input sentence;

$$\mathbf{x}' = \text{concatenate}(\mathbf{x}, \mathbf{c}_0, \cdots, \mathbf{c}_{N-1}). \tag{3.2}$$

  Then, the output of TwoC encoder is the output of a stacked transformer encoder with $\mathbf{x}'$. The computational complexity is $\mathcal{O}((L_s + NL_c)^2)$, where L_c is a fixed length of context sentences. The complexity becomes quadratically expensive as N grows.

- **Discourse Aware Transformer (DAT) [17]**: DAT handles context sentences with an extra context encoder which is also a stacked transformer encoder. We slightly modified DAT to make it available at handling multiple context sentences since [17] is originally designed for handling a single context sentence.

  The context encoder has the same structure and even shares its weights with the source encoder through $N_{Layer} - 1$ layers. In the last layer, the context encoder has another transformer encoder module without sharing its weights. The last layer of the source encoder takes an intermediate output tensor $\mathbf{h}'$ which

is resulted from $N_{Layer} - 1$ stacked transformer encoder, processes both self-attention and context-source attention with $\mathbf{t}$ using $\mathrm{MultiHead}$;

$$\begin{aligned}\mathbf{t} = \mathrm{concatenate}(&\mathrm{StackedTransformerEncoder}(\mathbf{c}_0), \cdots, \\ &\mathrm{StackedTransformerEncoder}(\mathbf{c}_N)),\end{aligned} \tag{3.3}$$

$$\mathbf{h}_{context} = \mathrm{MultiHead}(\mathbf{h}', \mathbf{t}, \mathbf{t}), \tag{3.4}$$

and

$$\mathbf{h}_{source} = \mathrm{MultiHead}(\mathbf{h}', \mathbf{h}', \mathbf{h}'). \tag{3.5}$$

the final output tensor of encoder h is given with the gated sum as follows;

$$\mathbf{h} = \sigma(\mathbf{W}_h[\mathbf{h}_{source}, \mathbf{h}_{context}] + \mathbf{b}_\mathbf{h}), \tag{3.6}$$

where $\mathbf{W}_h$ is a learnable weights and $\mathbf{b}_\mathbf{h}$ is a learnable bias term.

The computational complexity of DAT is $\mathcal{O}(L_s^2 + NL_c^2)$, which is comparable to our model. However, in order to process context-source attention with multiple context sentences, it concatenates all tensors from each context encoders to a long tensor where long-range dependencies of SAN may be limited.

- **Document-level Context Transformer (DCT) [55]**: The encoder of DCT is similar to the DAT, except for the integration of the context and source encoder. Instead of context-source attention and gated sum at the output of both encoders, each layer of the source encoder takes encoded contextual information t and compute context-source attention followed by point-wise feed-forward layer;

$$\mathbf{h}_{context} = \mathrm{MultiHead}(\mathbf{h}', \mathbf{t}, \mathbf{t}), \tag{3.7}$$

and

$$\mathbf{h} = FFN(\mathbf{h}_{context}). \tag{3.8}$$

Since the extensive use of the context-source attention in the encoder, the computational complexity of DCT is $\mathcal{O}(NL_cL_s + L_s^2 + NL_c^2)$. This can grow prohibitively, especially on handling long context sentences or when the number of context sentences is large.

- **Hierarchical Attention Networks (HAN) [22]**: HAN has a hierarchical structure with two stage at every HAN layer. At the first level of the hierarchy, a single HAN layer encodes each context sentence $\mathbf{c}_i$ to an intermediate tensor $\mathbf{e}_i \in \mathbb{R}^{L_c \times D}$ with context-source attention;

$$\mathbf{e}_i = \mathrm{MultiHead}(\mathbf{h}', \mathbf{c}_i, \mathbf{c}_i), \tag{3.9}$$

where $\mathbf{h}'$ denotes an output from a previous layer or an input $\mathbf{x}$. Each $\mathbf{e}_i$ is a tensor with a length of L_c and let $\mathbf{e}_i^j$ be the j-th vector of $\mathbf{e}_i$.

At the second level of hierarchy, $\mathbf{e}_i^j$ in all context sentences are concatenated through i dimension, resulting tensors $\mathbf{s}^j \in \mathbb{R}^{N \times D}$;

$$\mathbf{s}^j = \mathrm{concatenate}(\mathbf{e}_0^j, \cdots, \mathbf{e}_N^j), \tag{3.10}$$

where N is a number of context sentences. Then, an intermediate output tensor t which contains contextual information queried by each word from the input sentence can be given as follows;

$$\mathbf{t} = \mathrm{MultiHead}(\mathbf{h}', \mathbf{s}^j, \mathbf{s}^j). \tag{3.11}$$

All $\mathrm{MultiHead}$ layers are followed by position-wise feed forward layers and normalization layers. Finally, the output tensor $\mathbf{h}$ of HAN encoder is computed with a gated-sum module introduced by [59]. The aforementioned structure of a single layer in HAN is stacked N_{Layer} times.

The computational complexity of HAN encoder is $\mathcal{O}(NL_cL_s + L_s^2 + NL_c^2)$ which is also comparable to our proposed model. Nonetheless, HAN encoder requires context-source attention two times at every layers. Also, since the second context-source attention is performed on $\mathbf{s}_i = \mathrm{concatenate}(\mathbf{e}_0^j, \cdots, \mathbf{e}_N^j)$, HAN does not take account of internal correlations among $(\mathbf{e}_i^0, \cdots, \mathbf{e}_i^{L_c})$.

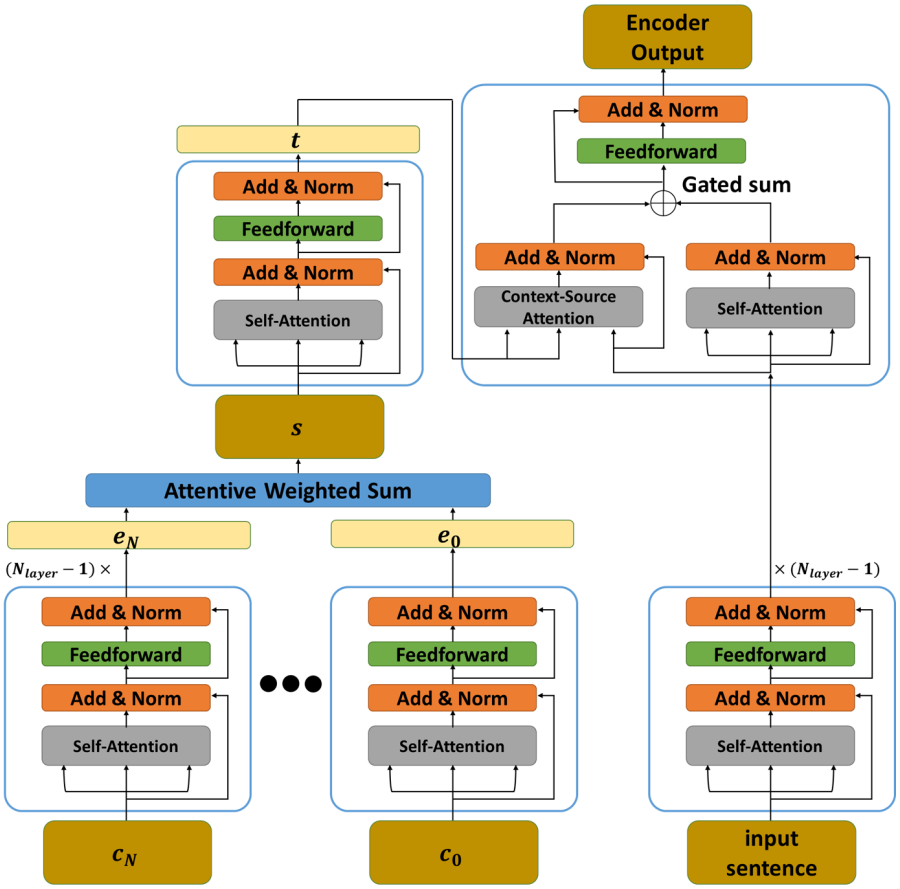

Figure 3.1: The structure of our proposed Hierarchical Context Encoder. Each context sentences $\mathbf{c}_i$ is encoded through transformer encoders to the tensor $\mathbf{e}_i$ and the attentive weighted sum module vectorizes each $\mathbf{e}_i$ to the vector $\mathbf{s}_i$. Upper transformer encoder encodes the input tensor $\mathbf{s}$ composed by concatenation $\mathbf{s} = (\mathbf{s}_0, \cdots, \mathbf{s}_N)$ and outputs our final context representation tensor $\mathbf{t}$. Then the context representation is combined to the source encoder by gated sum.

3.2.2 Hierarchical context encoder

We propose a novel context encoder that hierarchically encodes multiple sentences into a tensor. Our proposed encoder, Hierarchical Context Encoder (HCE), is designed to capture correlations between sentences in contexts as well as correlations between words in each sentence.

Each context sentence $\mathbf{c}_i$ after word embedding layer is given as a tensor of order 2; $\mathbf{c}_i \in \mathbb{R}^{L_c \times D'}$ where L_c is a maximum length of each context sentence and D' is a dimension of word embeddeding vectors. In the lower part of hierarchy, HCE encodes each of $\mathbf{c}_i$ to sentence-level tensor $\mathbf{e}_i$ using the stacked transformer encoder as [7];

$$\mathbf{e}_i = \mathrm{StackedTransformerEncoder}(\mathbf{c}_i). \tag{3.12}$$

Each encoded sentence-level tensor $\mathbf{e}_i$ is also a tensor of order 2, $\mathbf{e}_i \in \mathbb{R}^{L_c \times D}$ where D is a hidden dimension.

We then compress each encoded sentence-level tensor into a sentence-level vector by a self-attentive weighted sum module which is similar to that of [60]. Our self-attentive weighted sum module takes $\mathbf{e}_i$ as an input tensor and computes a vector $\mathbf{s}_i$ as follows;

$$\mathbf{s}_i = \sum_j \alpha_j \mathbf{e}_{ij}, \tag{3.13}$$

$$\alpha = \mathrm{FFN}(\mathrm{MultiHead}(\mathbf{e}_i, \mathbf{e}_i, \mathbf{e}_i)). \tag{3.14}$$

The output of the attentive weighted sum module $\mathbf{s}_i$ is a vector representing the information of each i-th context sentence. Then we concatenate $(\mathbf{s}_0, \cdots, \mathbf{s}_N)$ to a context embedding tensor $\mathbf{s}$. The context embedding tensor $\mathbf{s} \in \mathbb{R}^{N \times D}$ is fed into another SAN encoder layer which is the upper part of the hierarchy to encode the whole contextual information into a single tensor $\mathbf{t}$;

$$\mathbf{t} = \mathrm{TransformerEncoder}(\mathbf{s}). \tag{3.15}$$

Finally, the contextual information tensor t is combined to source encoder by gated

sum as Equation 3.4, 3.5, and 3.6, which is the same process introduced by [17]. Full structure of HCE is depicted in Figure 3.1.

The main difference between HCE and other baseline models especially HAN is that HCE encodes each context sentence as the way of sentence embedding with self-attention independent to the source word, while HAN uses context-source attention. To explain more in detail, two main differences between the hierarchical SAN structures of HAN and HCE are as follows: 1) at the bottom part of the hierarchy, HCE encodes each context sentence to a tensor with self-attention while HAN encodes each context sentence with context-source attention using query words from input sentences; and 2) at the upper part of the hierarchy, HCE first uses the self-attentive weighted sum to encode a tensor into a vector which contains the whole information from each context sentence, then encodes the whole contexts with self-attention again. On the other hand, HAN uses context-source attention again. To summarize, HCE only models the context-source relations at the upper part of the hierarchy resulting in a simpler and clearer model structure.

The computational complexity of HCE is $\mathcal{O}(L_s^2 + NL_c^2)$. HCE extracts more compact context-level representation from each sentence-level representation by self-attentive weighted sum over each e_i, hence it complements DAT [17] and DCT [55] whereas they take the whole contexts as a single sentence by concatenation. Besides, the encoding procedure of context sentences is not dependent on the input sentence x unlike HAN. This allows HCE to cache context-level representations t of frequently appeared context sentences, which is important in implementing a real-time application.

3.3 Data

We experimented with our model and baseline models on English-German TED corpus, English-German OpenSubtitles corpus, English-Turkish OpenSubtitles corpus,

and our web-crawled English-Korean subtitle corpus.

3.3.1 English-German IWSLT 2017 corpus

We use the English-German corpus from the IWSLT 2017 evaluation campaign [61], which is publicly available on WIT^3 website[1]. The corpus consist of transcriptions and their translations of TED talks. We combine `dev2010` and `tst2010` into a development(*dev*) set and `tst2015` as a *test* set. We extract context-aware dataset where each set consists of a *source*, a *target* sentence and multiple *context* sentences. Since the corpus is aligned as sentence level, we assume that every 2 preceding sentences are *context* sentences. We also include context sentences only within the same talk of the source sentence, as the data is separated as talks. The resulting dataset consists of $211k$, $2.4k$, $1.1k$ examples of *train*, *dev*, *test* sets respectively. Also, we put a special *beginning of context* token at the beginning of each context sentences to differentiate from source sentences. Finally, we have used a byte-pair encoded vocabulary with about 16,000 tokens.

3.3.2 OpenSubtitles corpus

We also choose the OpenSubtitles corpus for English-German and English-Turkish tasks. We use the 2018 version [62] of the data, each consist of $24.4M$, $47.4M$ parallel sentences respectively. Following the approach in [21], we first cleaned the data by picking only pairs with a time overlap of subtitle frames at least 0.9. After cleaning, we take $7.5M$ and $20.2M$ sentences for English-German and English-Turkish corpus.

We then take the *context* sentences by using the timestamp of each subtitle. The timestamps contain start time and end time in ms for each subtitle. We focus on the start times to compile a set of data including a source sentence and preceding contextual sentences. We assume that if the start time of a preceding sentence is within 3000 ms from the start time of a sentence then that preceding sentence contains the contex-

[1] https://wit3.fbk.eu/mt.php?release=2017-01-trnted

Start Time	End Time	English	Korean
		...	
337733	339967	Daniel likes hanging out with his cousins.	다니엘은 사촌들과 노는걸 좋아했거든요
340035	341168	He's been going back and forth until Leith and I	양육권을 제대로 가질 수 있을때까지
341236	342303	can settle custody.	왔다 갔다 했어요
344373	345940	Listen, don't worry.	너무 걱정 마세요
		...	

Figure 3.2: Bilingual subtitle samples from our English-Korean test files

tual information. We set the maximum number of preceding contextual sentences up to 2.

3.3.3 English-Korean subtitle corpus

Finally, for English-Korean experiments, we construct a web-crawled subtitle corpus with 5,917 files. These files are English-Korean bilingual subtitle files of movies, TV series, and documentary films from various online sources. We set randomly selected $5.3k$ files for *train*, 500 files for *dev*, and 50 files for *test* set. The *train* set includes $3.0M$ sentences, the *dev* set includes $28.8k$ sentences, and the *test* set includes $31.1k$ sentences. Our web-crawled English-Korean bilingual subtitle files include time stamps for each subtitles. Thus we pre-process those files as similar as processing in Section 3.3.2. The resulting data have $1.6M$ sets of serial sentences in *train set*, $155.6k$ sets in *dev set*, and $18.1k$ sets in *test set*. We also have used a byte-pair encoded vocabulary with about 16,500 tokens for English-Korean experiments. We display some raw samples from our test files in Figure 3.2.

3.4 Experiments

We evaluate our HCE by BLEU score, model complexity, BLEU on helpful/unhelpful set, and accuracy on the pronoun resolution set. All experimental results show the effectiveness of HCE compared to baseline models.

3.4.1 Hyperparameters and Training details

Through our experiments, we use 512 hidden dimensions for all layers including words embedding layers, SAN layers, and the encoded context layer. We set $N_{Layer} = 6$ for all models and share the weights of the source encoder to context encoder for the DAT, HAN, and HCE models. For all attention mechanisms, we set the number of heads as 8. The dropout rate of each SAN layers is set to 0.1.

For each language pair, we tokenize each text by the wordpiece model [63, 8] with a vocabulary of about 16,000 tokens. Also, we put a special *beginning of context* token <BOS> at the beginning of each context sentences to differentiate from source sentences.

We implement all the evaluated models using the `tensor2tensor` framework [64]. We train all models with ADAM [65] optimizer with learning rate 1e-3 and adopt early stopping with *dev* loss. Unlike [22, 55, 23], we do not use the iterative training which trains the model on a sentence-level task first, then fine-tunes the model with contextual information. All the models we have evaluated are trained from scratch with random initialization.

For scoring BLEU, we use the `t2t-bleu` script[2] which outputs the identical results as Moses script [66].

3.4.2 Overall BLEU evaluation

We measure performances of HCE and other five baseline models in English-German (IWSLT'17 and OpenSubtitles), English-Turkish (OpenSubtitles), and English-Korean(our Web-crawled corpus). Overall BLEU scores on all eight datasets are displayed in Table 3.1. Our model yields the best performances on all eight datasets. Especially on our Web-crawled English-Korean, HCE shows superior performance compared to other models. These results indicate that our model exploits given contextual sentences effectively and translate better than all five baseline models in English-German, English-Turkish and English-Korean translation tasks.

3.4.3 Model complexity analysis

We also observe that our HCE is the most efficient in training speed and inference time among our baselines. In Table 3.2, HCE records the fastest training speed and inference time indicating that HCE has the most computationally efficient structure. These

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

Corpus	IWSLT'17		OpenSubtitles		OpenSubtitles		Web-crawled	
Language pair	En→De	De→En	En→De	De→En	En→Tr	Tr→En	En→Ko	Ko→En
Transformer without contexts	28.25	32.18	27.95	33.93	24.89	36.27	8.58	23.67
Transformer with contexts	28.65	32.68	28.07	34.04	23.96	35.81	9.46	24.23
DCT [55]	26.76	30.33	26.3	32.05	21.91	34.3	6.5	20.72
DAT [17]	28.82	32.59	28.09	33.99	24.30	35.23	8.56	23.91
HAN [22]	28.85	32.72	28.00	34.42	24.86	36.55	8.76	24.41
HCE (ours)	**28.89**	**33.01**	**28.40**	**34.59**	**25.11**	**36.84**	**11.30**	**26.70**

Table 3.1: BLEU score. Our proposed Hierarchical Context Encoder have shown the best results in all language pairs.

Model	Training speed (steps/sec)	Inference time (tokens/sec)	# of Params
TwC	4.07	62.10	61.0M
DCT	2.42	45.32	98.7M
DAT	4.59	65.07	69.9M
HAN	4.47	64.05	66.2M
HCE	**4.67**	**65.12**	66.7M

Table 3.2: Training speed, inference time and number of parameters.

results also show that the performance gain of HCE is not only from the complexity of the model but the structural strength because the number of parameters is comparable to others.

3.4.4 BLEU evaluation on helpful/unhelpful context

Model	Total set	helpful set	unhelpful set	BLEU gap
Transformer without contexts	7.46	6.69	8.04	+1.35
Transformer with contexts	8.29	7.45	8.92	+1.47
DAT [17]	8.22	7.48	8.77	+1.29
HAN [22]	8.34	7.44	9.01	+1.57
HCE (ours)	10.27	10.08	10.40	**+0.32**

Table 3.3: BLEU score evaluations with helpful contexts set and unhelpful contexts set from En→Ko test data. All four baseline models have shown large gap between BLEU score on *helpful* contexts set and BLEU score on *unhelpful* contexts set. On the other hand, Our proposed Hierarchical Context Encoder has almost closed the gap between BLEU scores on two sets.

In order to verify that our model actually uses the contextual information to improve translation quality, we conduct an additional experiment with a part of data

where contextual sentences are helpful for translating and the other part of data where they are not. We randomly choose 10,000 sets of serial sentences from our *test set* of En→Ko data and split them up into two parts by crowd-sourcing with Amazon Mechanical Turk [67]. The first part consists of 4,331 sets of which context sentences are helpful for translating (*e.g.* context sentences include critical information, exact referred object by pronouns, or residual parts of an incomplete source sentence). The remaining part consists of 5,669 sets of which context sentences are unrelated to translate the source sentences.

We examine BLEU scores of two parts separately to observe how well each model uses helpful contexts. The results are displayed in Table 3.3. We observe a large gap between BLEU score on *helpful set* and that on *unhelpful set* with all four baseline models, showing that *helpful set* is harder to translate because abstracting and exploiting contextual information is likely to be mandatory to translate *helpful set*. On the other hand, HCE closes the gap between BLEU scores on each set, indicating that HCE understands the contextual information and is able to perform on *helpful set* as well as on *unhelpful set*.

3.4.5 En→Ko pronoun resolution test suite

Finally, we evaluate the accuracy of all models that use contexts on our En→Ko pronoun resolution test suite. we create a test suite for English→Korean pronoun resolution to examine how well a model understands contextual information. Our test suite is composed of 150 sets, each of which includes 1) a source sentence with a pronoun, 2) preceding contextual sentences with the exact word referred to by the pronoun, 3) a target sentence with the corresponding pronoun, 4) a *correct* target sentence where the pronoun is replaced with the exact word, and 5) a *wrong* target sentence where the pronoun is replaced with an unrelated word. We follow a scoring method in [1] for evaluation; if a model's negative log-likelihood of *correct* sentence is lower than that of *wrong* sentence, then we consider the model is able to detect wrong pronoun

Label	English	Korean
context 1	When did the **tower** collapse?	
context 0	Oh, last winter.	
source / target	Brother Remigius says we haven't the funds to repair **it**.	레미져스 수사님 말론 **그걸** 고칠 돈이 없다는군.
correct		레미져스 수사님 말론 **탑**을 고칠 돈이 없다는군.
wrong		레미져스 수사님 말론 **지붕**을 고칠 돈이 없다는군.

Figure 3.3: A sample set of English→Korean pronoun resolution test suite

Model	accuracy
Transformer with contexts	0.25
DAT [17]	0.44
HAN [22]	0.47
HCE (ours)	**0.48**

Table 3.4: Accuracy on our En→Ko pronoun resolution test suite.

translation.

A sample from our test suite is displayed in Table 3.3, the pronoun and corresponding words are emphasized in bold. In the sample, the *source* sentence has a pronoun "**it**" referring the word "**tower**" in the *context 1* sentence. The *target* sentence also has the corresponding boldfaced pronoun in Korean, "그걸 (it)". We replace the pronoun in *target* sentence to the exact referring Korean word "탑 (tower)" in the *correct* sentence, and we replace it to an unprecedented yet similar Korean word, "지붕 (roof)" in the *wrong* sentence.

The results are displayed in Table 3.4. While TwC scores the lowest accuracy with 0.25, DAT and HAN record accuracy with 0.44 and 0.47 respectively. HCE records the highest accuracy of 0.48 in this test. These results support the hypothesis that it is harder to capture contextual information on a single long concatenated sentence than on structured multiple context sentences. Also, the result that HCE and HAN both perform better than DAT reveals the strength of hierarchical structure for multiple contexts which is able to capture the contextual information effectively.

3.4.6 Qualitative Analysis

Attention Visualizations

Figure 3.4 shows three examples how contextual encoders attend and comprehend the context sentences while translating a particular pronoun. The words in brackets next to the input sentences are the words in context sentences referred by each boldfaced

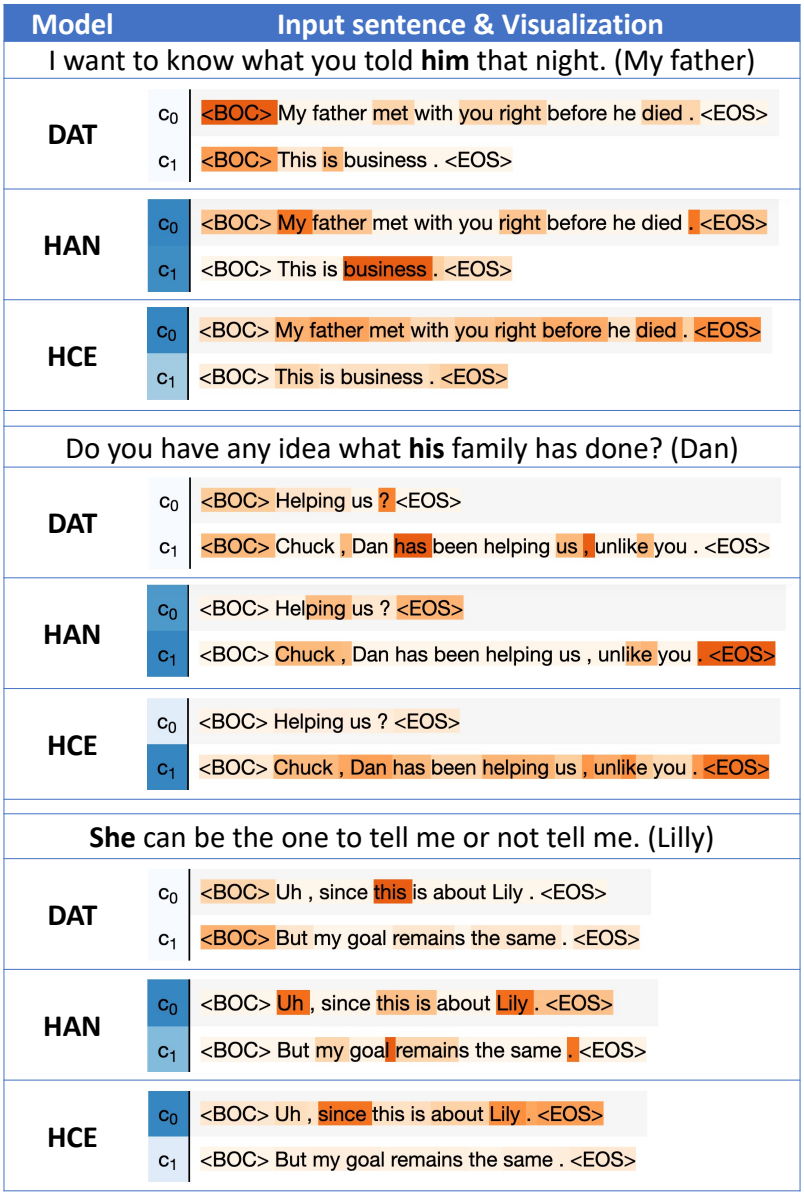

Figure 3.4: Three visualization examples of attention weights for given pronoun bold-faced words which are referring to the words in brackets. We refer each of them as **(a)** the uppermost example, **(b)** the middle example, and **(c)** the bottom example.

pronoun. The intensity of color (orange) is proportional to the attention weight for each word. Also, the intensity of color (blue) is proportional to the attention weight for each context sentence in HCE and HAN.

In general, the third-person pronouns in English are often translated into Korean pronouns that do not contain attributes like gender, or phrases indicating the referenced person or object. For example, the word "his" in the middle example (b) has translated as "제네 (their)" which is a correct Korean possessive pronoun for referring "Dan" in c_1 sentence. In the bottom example (c), the word "She" has translated as "본인 (oneself)" which can be used for both male and female. Likewise, the word "him" in the uppermost example (a) has translated as "아버지 (father)" which is the exact referred word. Considering such phenomena, we regard that correctly referencing the proper nouns is crucial in translating pronouns into Korean.

From this point of view, Table 3.4 explains the strength of HCE in the En→Ko translation. As presented in Table 3.4, we observed that HCE gives more attention to the context sentences which contain the exact referred words. Hence, the upper hierarchy of HCE pays its attention to the more important sentence as we have intended. We also observed that both our HCE and HAN tend to attend to nouns such as names of people (e.g. Dan, Chuck) or names of specific locations (e.g. the church, Paris). Nevertheless, HCE more accurately attends to the exact referred words comparing to HAN. In the first example, HCE gives large portion of its attention to "My father" while HAN choose "business" as the most important word. The second example also shows the ability of HCE to exploit context information properly. HCE understands that the word "Dan" is more important than "Chuck", while HAN gives most of its attention to the word "Chuck" except for the <EOS> token. Although HCE computes context representations independent of the input query, these visualization examples show that HCE can correctly attend to the exact words referred by the pronouns.

Source	Why would he pay you $10 million / **not to ask any of these questions?**
Reference	왜 아버지는 아무 조건도 없이 / **당신에게 천만 달러를 줬을까요?**
TwoC	이런 질문은 하지 말라는 건가요? (Not to ask this kind of question?)
HCE	왜 당신에게 돈을 지불한 거죠? (Why would him/her pay you the money?)

(a)

Context	This country does not negotiate with terrorists. Tell that to the families o f the dead!
Source	Do you have any idea what the public reaction will be when word gets o ut that these are acts of terrorism?
Reference	테러로 인한 참사인 것이 알려지면 대중의 반응이 어떨지 아십니까?
HAN	테러 행위가 밝혀지면 대중의 반응이 어떨지 아십니까? (If a/another terrorism act is revealed)
HCE	테러 행위로 알려지면 대중의 반응이 어떨지 아십니까? (If it is revealed as a terrorism act)

(b)

Context	Man asked us to locate his daughter.
Source	It was a long time ago, but it looks like she might have known Dewall.
Reference	오래전에 실종됐지만 그 애가 드웰을 아는것 같아서요.
HCE	오래전 일이지만, 그녀가 데웰을 알고 있는 것 같아요.
Google Translate	오래 전 일이지만 그녀는 Dewall을 알고 있었던 것 같습니다.
Naver Papago	오래 전 일이지만 뎁을 알지도 몰라.

(c)

Figure 3.5: Translation samples. Context, source, and reference sentences are from our En-Ko test set. **(a)** Translations of a sentence fragmented into the context and the source. Each model is given with a context sentence (unbolded) and a source sentence (bolded). **(b)** Considering the context, HCE's translation is more adequate than HAN. **(c)** Comparison with two commercial MT systems, Google Translate and Naver Papago that are retrieved on 15 Dec 2021 by providing a concatenation of context and source sentences as input.

Translation Samples

Figure 3.5 shows 3 sets of source, context, and reference target sentences from our En-Ko test sets and their translations. Fig. 3.5-(a) displays an example of fragmented source sentence and its translations. In the example, the English sentence "Why would he pay you $10 million not to ask any of these questions?" is divided into two pieces, inputted as a context and a source sentence respectively. This situation occurs very frequently in subtitle translations and poses a challenge to MT models combined with different word orders of both languages (En: SVO vs Ko: SOV for example). Despite the difficulties, the HCE can able to generate translations of the relevant part of the source sentence compared to the context-agnostic model (TwoC).

We also compare the HCE with HAN, the best performing context-aware baseline model in Fig. 3.5-(b). Although the Korean translations differ in just one morpheme ('가' vs '로'), their semantics are completely different. The context sentence suggests that the speaker of the source sentence is afraid that the incident would be revealed as a terrorist act. Considering the context, HCE's translation is more adequate than HAN, reflecting the strength of HCE in incorporating contextual information.

Finally, we compared HCE with Google Translate and Naver Papago in Fig. 3.5-(c) which are two of the commonly used commercial MT services. In these examples, HCE's translations are adequate with respect to both the source and reference sentences and also comparable with the two commercial systems. In fact, both HCE and Naver Papago have generated more fluent Korean sentences than Google since the names of people are also written in Korean despite minor spelling differences.

On the other hand, HCE shows some shortcomings commonly found in NMT models. In the example in figure 3.6-(a), HCE has omitted the translation of the conjunction word "*Then*", while Google and Naver results contain corresponding translations. It shows the phenomenon known as word omission [68]. Fig. 3.6-(b) displays another translation error that the HCE did not translate the source sentence at all. Instead, HCE just copied the source English sentence. This is due to the fact that the training data

Context	I called my wife, told her I wanted to take a break.
Source	Then at Erik Dalton's party, Cole starts laughing.
Reference	그런데 에릭 달튼의 파티에서 콜이 갑자기 웃기 시작하는 거에요.
HCE	에릭 달튼의 파티에서 콜이 웃음을 시작했죠.
Google Translate	그런 다음 Erik Dalton의 파티에서 Cole은 웃기 시작합니다.
Naver Papago	그리고 에릭 달튼의 파티에서 콜은 웃기 시작했다.

(a)

Context	♪ hanging at the bar ♪
Source	♪ I don't owe you a thang, baby ♪
Reference (as in test set)	♪ I don't owe you a thang, baby ♪
HCE	♪ i don't owe you a thang, baby ♪
Google Translate	♪ 난 너에게 빚진 게 없어, 자기 ♪
Naver Papago	♪ 네게 땡땡이를 빚진 게 아니야, 자기야 ♪

(b)

Figure 3.6: Unsuccessful translation samples. Google and Naver translations are retrieved on 15 Dec 2021. **(a)** HCE omitted the translation of ”Then” compared to Google Translate and Naver Papago. **(b)** HCE just copied the source sentence as the train/test data, but this is an unexpected behavior as a general MT system.

is composed of the subtitle (and also is the test data). As can be seen from the special character denoting a musical note, the source sentence in 3.6-(b) is a part of song lyrics in the video. Since many subtitle creators tend to just copy the source lyrics rather than translate them, our En-Ko subtitles data contains such untranslated lyrics with musical notes. Copying the English lyrics in the example can be explained in this way since the model is trained to mimic this behavior as seen in the data (as a form of exposure bias). However, this is a completely unexpected outcome as a general MT system as Google and Naver are able to translate the lyrics into Korean. We suggest that these limitations may be due to the fact that HCE does not employ any special techniques to deal with these phenomenons and need to be enhanced for practical applications.

3.5 Summary of Efficient Hierarchical Architecture for Modeling Contextual Sentences

In this chapter, we have introduced Hierarchical Context Encoder (HCE) structure which is able to encode multiple contextual sentences with hierarchical SAN structure. We have shown that our model outperforms all baseline models in English-German, English-Turkish and English-Korean translation tasks and also that our model is the most efficient in computational complexity. We also have shown that our model closes the gap of translation quality between the sentences with helpful contexts and the sentences with unrelated contexts, indicating that our model is better at exploiting the helpful contextual information for translating than baseline models. Analysis on pronoun resolution test suite support the effectiveness of our HCE.

Chapter 4

Contrastive Learning for Context-aware Neural Machine Translation

Neural machine translation (NMT) has achieved impressive performances on translation quality, due to the introduction of novel deep neural network (DNN) architectures such as encoder-decoder model [5, 6], and self-attentional networks (SANs) like Transformer [7]. The state-of-the-art NMT systems are now even comparable with human translators in sentence-level performance.

However, there are a number of issues on document-level translation [69]. These include pronoun resolution across sentences [70], which needs cross-sentential contexts. To incorporate such document-level contextual information, several methods for context-aware NMT (CNMT) have been recently proposed. Many of the works have focused on introducing new model architectures like multi-encoder models [17] for encompassing contextual texts of the source language. These works have shown significant improvement in addressing discourse phenomena such as anaphora resolution mentioned above, as well as moderate improvements in overall translation quality [71].

Despite some promising results, most of the existing works have trained the model by minimizing cross-entropy loss, making the model rather exploit contextual information implicitly such as a form of regularization [25, 26]. Data augmentation for CNMT

is also an important issue, despite that recent works have focused on back-translation [72].

In this chapter, we propose a Coreference-based Contrastive Learning for context-aware NMT (CorefCL), a novel data augmentation and contrastive learning scheme leveraging coreference information. Cross-sentential coreference between the source and target sentence can be a good source of training signal for CNMT since it occurs when one or more expressions refer to the same entity, thus reflects dependencies between the source and contextual sentences.

CorefCL starts by conducting automatic annotation of coreference between the source and contextual sentences. Then, the referred mentions on contextual sentences are corrupted by removing and/or replacing tokens to generate contrastive examples. With those contrastive examples, we introduce a contrastive learning scheme equipped with a max-margin loss which encourages the model to discriminate between the original examples and the contrastive ones. By doing so, CorefCL makes the model more sensitive to cross-sentential contextual information.

We experimented with CorefCL on three English-German corpora and one English-Korean document-level corpus, including WMT, IWSLT TED talk, and OpenSubtitles'18 English-German subtitles translation task, and a web-crawled English-Korean subtitles translation. In all translation tasks, CorefCL consistently improves overall BLEU over vanilla CNMT models. On experiments with three common context-aware model settings, we show that improvements by CorefCL are also model-agnostic. Finally, we show that the proposed method significantly improved the performance on ContraPro [1], an English-German contrastive coreference benchmark.

4.1 Related Works

4.1.1 Context-aware NMT Architectures

CNMT has been vigorously studied to exploit the crucial context information in surrounding sentences. Recent works have shown that contextual information can help the model to generate not only more consistent but also more accurate translation [1, 13, 17, 25].

In particular, [17] introduced a context-aware Transformer model which is able to induce anaphora relations, [22] showed that a model using cross-sentential contextual information significantly outperforms in document-level translation tasks, and [24] insisted that context-aware models record the best performance especially in spoken language translation tasks where mandatory information tend to be sparse over multiple sentences.

The simplest method for CNMT is to concatenate all surrounding sentences and treat the concatenated sequence as a single sentence [18]. Although the concatenation strategy boosted translation quality of SAN-based architectures in multiple tasks, it lagged behind efficiency as the SAN has limited long-range dependency [56].

To improve the efficiency, an additional encoder module is introduced to encode only the context sentences [14, 17, 73]. Additionally, hierarchical structures also have been introduced because the context sentences do not have the same significance as the input sentences [22, 24].

For training CNMT, most of existing studies relied on conventional negative log-likelihood (NLL) minimization similar to the sentence-level systems. Since this do not directly uses contextual information, several methods have been proposed to complement the insufficiency, i.e. adding context-dependent regularization [75], introducing reinforcement learning [76], or curriculum learning [77]. Inspired by these approaches, we introduce contrastive learning that exploits contextual dependency among source sentences.

4.1.2 Coreference and NMT

The difference in coreference expressions among languages [74, 78] gives MT systems a challenge on pronoun translation [19]. Several recent works have attempted to incorporate coreference information [79]. The closest work to ours is [80] which also adds noise on creating a coreference-augmented dataset, while we do not add oracle coreference information directly to the training data.

4.1.3 Data augmentation for NMT

One of the most common methods for data augmentation in NMT is back-translation that generates pseudo-parallel data from monolingual corpora using intermediate NMT models [37]. Generally, back-translation is conducted at sentence-level, however, several works have proposed document-level back-translation [72, 81].

On the other hand, sentence corruption by removing or replacing word(s) has also been widely used for improving model performance and robustness [21, 82]. Inspired by these works, we choose sentence corruption for contrastive learning.

4.1.4 Contrastive Learning

Contrastive learning is to learn a representation by contrasting positive and negative (contrastive) examples. It has succeed in various machine learning fields including computer vision [83] and natural language processing tasks like word [84] and sentence representation learning [85], as well as sequence-to-sequence learning [86].

Recently, several approaches to contrastive learning for NMT have also been studied. Yang et. al. [68] proposed strategies for generating word-omitted contrastive examples and leveraging contrastive learning for reducing word omission errors in NMT. Pan et. al. [87] applied contrastive learning for multilingual MT and employed data augmentation for obtaining both the positive and negative training examples.

While these works have been conducted in sentence-level NMT settings, we focus on extending contrastive learning in context-aware NMT.

4.2 Context-aware NMT models

In this section, we briefly overview context-aware NMT methods and describe our baseline models which are also commonly adopted in recent works.

Generally, a sentence-level (context-agnostic) NMT model takes an input sentence in a source language and returns an output sentence in a target language. On the other hand, a context-aware NMT model is designed to handle surrounding contextual sentences of source and/or target sentences. We focus on leveraging the contextual sentences of the source language.

Throughout this work, we consider self-attentional networks (SANs) like Transformer [7] by following the majority of the recent works on context-aware NMT. We list four SAN-based configurations that we used in the experiments:

- **sent-level, sent-level-t5**: Vanilla sentence-level SAN as same settings as the Transformer, that ignores contextual sentences. In addition to **sent-level** which is a SAN trained from scratch, we also experimented with a pre-trained language model (PLM) fine-tuned for English-German task. In our settings, we use the Text-to-Text Transfer Transformer (T5) [53] refered as **sent-level-t5**.

- **concat, concat-t5**: SAN with a concatenation of the input andits contextual sentences as an input [18]. Since this can incorporate contextual sentences without modifying the SAN model, we also implemented this settings in T5 as **concat-t5**.

- **multi-enc**: This has an extra encoder for encoding contextual sentences separately. We experimented with the Discourse Aware Transformer (DAT) [17].

- **multi-enc-hier**: Multi-encoder model with hierarchically computing contextual representations in token-level first, then sentence-level. We experimented with the HCE [24] introduced in the Chapter 3.

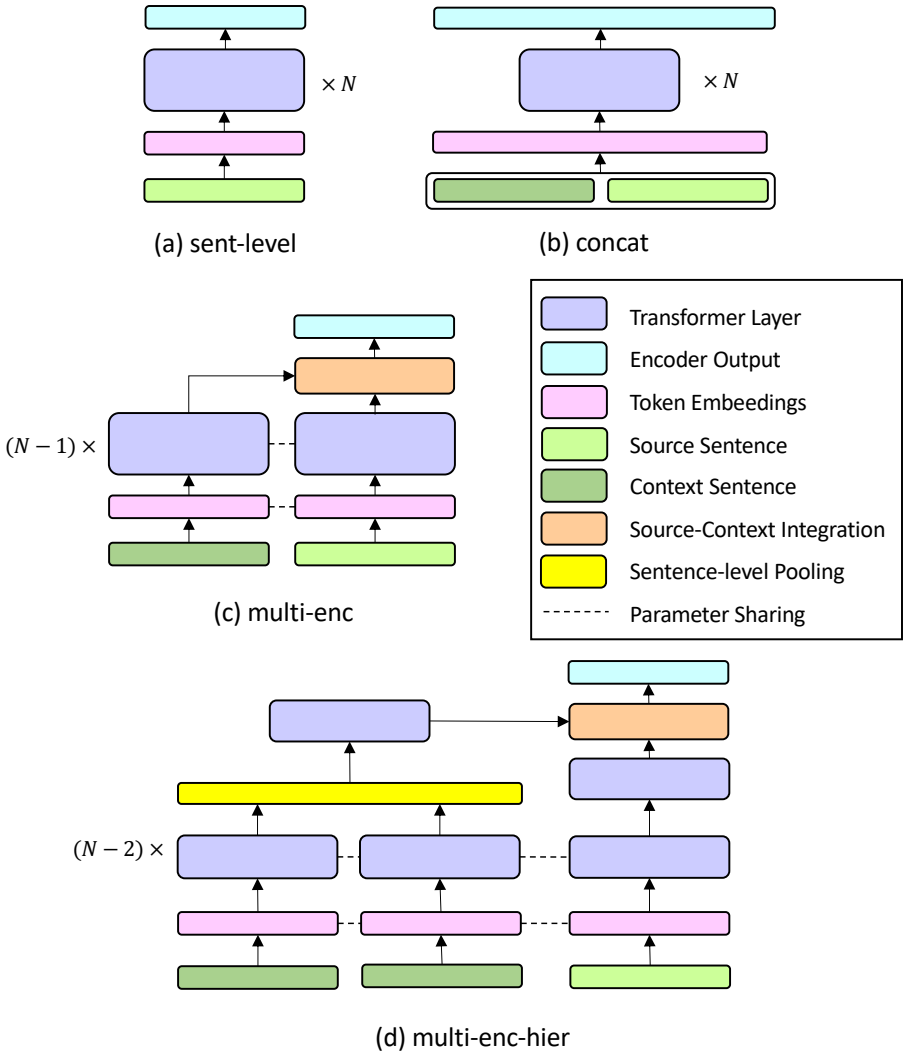

Figure 4.1: The structure of compared context-aware NMT models.

All the model structures are displayed in Figure 5.3. For detailed explanations on DAT and HCE, please refer to the Section 3.2.1

4.3 Our Method: CorefCL

In this section, we explain the main idea of CorefCL, a data augmentation and contrastive learning scheme leveraging coreference between the source and contextual sentences.

4.3.1 Data Augmentation Using Coreference

Generally, constrastive learning encourages a model to discriminate ground-truth and contrastive (negative) examples. In existing works, a number of approaches have been studied for obtaining contrastive examples:

- Corrupting the sentence by randomly removing or replacing one or more tokens in the sentence. [68]
- Choosing an irrelevant example in the batch or dataset. [87]
- Perturbations on representation space. Usually output vector of encoder or decoder is used. [86]

CorefCL basically takes a similar approach to the first one, by the sentence corruption. However, unlike previous works that modify the source sentence, CorefCL modifies the contextual sentences to form contrastive examples. Specifically, we corrupt cross-sentential coreference mentions which occur between the source and its contextual sentences. This is based on the intuition that coreference is one of the core components of coherent translation.

More formally, steps to forming contrastive examples in CorefCL are as follows (see also Figure 4.2):

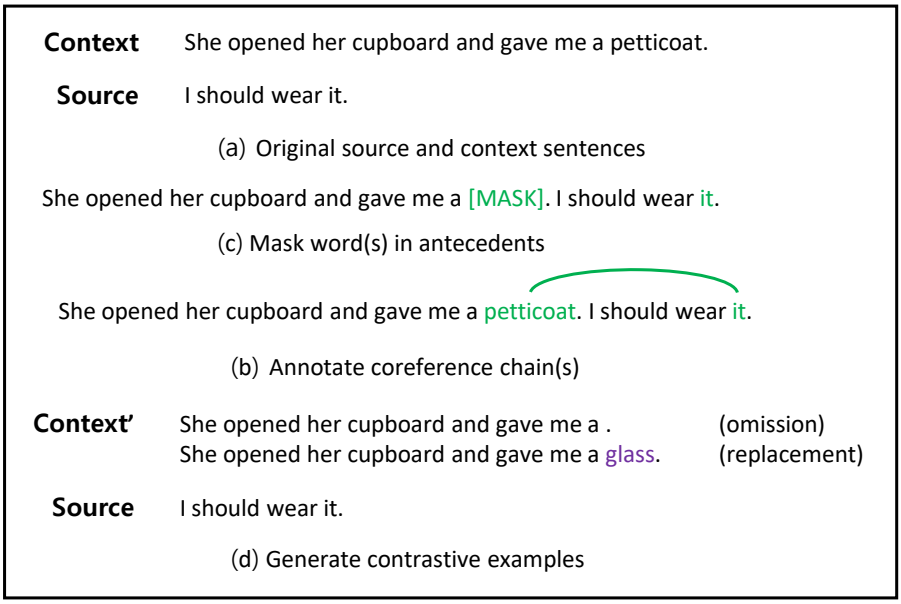

Figure 4.2: Data augmentation process of CorefCL.

1. Annotate the source documents automatically. We use NeuralCoref[1] to identify the coreference mentions between the source and its previous sentences as contextual sentences

2. Filter the examples with cross-sentential coreference chain(s) between the source and contextual sentences. Around 20 to 30% of the training corpus is annotated in this way. See Section 4.4.1 for details

3. For each coreference chain, mask every word in the antecedents with a special token. We also keep the original examples for training

4. Masked words are replaced randomly with other words in vocabulary (*word replacement*), or omitted (*word omission*)

In the experiments, we take both of the corruption strategies. Precisely, the masked words are removed with a probability of 0.5, or randomly replaced otherwise. We

[1] https://github.com/huggingface/neuralcoref

found that this method is more effective compared to the methods using only one of the two corruption strategies. Please refer to the ablation study in Section 4.4.5 for more details.

4.3.2 Contrastive Learning for Context-aware NMT

Context-aware NMT models can implicitly capture dependencies between the source and contextual sentences. CorefCL introduces a max-margin contrastive learning loss to train the model to explicitly discriminate inconsistent contexts. This contrastive loss also encourages a model to be more sensitive to the contents of contextual sentences.

Formally, given the source $\mathbf{x}$, target $\mathbf{y}$, n contextual sentences $C = [\mathbf{c}_1, \cdots, \mathbf{c}_n]$ in the data $\mathcal{D}$, we first train the model by minimizing a negative log-likelihood loss, which is a common MT loss:

$$\mathcal{L}_{MT} = \sum_{(\mathbf{x},\mathbf{y},C)\in\mathcal{D}} -\log P(\mathbf{y}|\mathbf{x}, C). \tag{4.1}$$

Once the model is trained with MT loss, we fine-tune the model with a contrastive loss. With a contrastive version of context $\tilde{C}$, our contrastive learning objective is minimizing a max-margin loss [68, 88]:

$$\mathcal{L}_{CL} = \sum_{(\mathbf{x},\mathbf{y},C,\tilde{C})\in\mathcal{D}} \max\{\eta + \log P(\mathbf{y}|\mathbf{x}, \tilde{C}) - \log P(\mathbf{y}|\mathbf{x}, C), 0\}. \tag{4.2}$$

Minimizing $\mathcal{L}_{CL}$ encourages the log-likelihood of the ground-truth to be at least η larger than that of the contrastive examples. In our formulation, we want the model to be more sensitive to the subtle changes in the contextual sentences.

The contrastive loss is jointly optimized with MT loss since we empirically found that the joint optimization has yielded better performance than minimizing CL loss only as similar to [89]:

$$\mathcal{L} = (1 - \alpha)\mathcal{L}_{MT} + \alpha\mathcal{L}_{CL},$$

where $\alpha \in [0, 1]$ is a weight for balancing between contrastive learning and MT loss. For simplicity, we fixed α during fine-tuning.

4.4 Experiments

4.4.1 Datasets

We experimented with CorefCL on various document-level parallel datasets: i) 3 English-German datasets including WMT document-level news translation[2] [90], IWSLT TED talk [3] [61], OpenSubtitles'18[4] [62], and ii) our web-crawled English-Korean subtitles corpus.

For all tasks, we take every 2 preceding sentences as contextual sentences and we only consider sentences within the same document (article, talk, movie, one episode of TV programs, etc.) of the source sentence. If split of the validation and the test set is not presented in the data, we apply document-based split to ensure that training and validation/test data is well-separated. Details of datasets are listed as follows:

WMT We use a set of parallel corpora annotated with document boundaries which is released in WMT'19 news translation task. Specifically, we combine Europarl v9, News Commentary v14, and MODEL-RAPID to form a training set containing $3.7M$ examples and $0.85M$ with cross-sentential coreferences. For validation and test sets, we used newstest2013 and newstest2019 which contain $3.05k$ and $2.14k$ examples respectively.

IWSLT The IWSLT dataset consists of transcriptions of TED talks in a variety of languages. We used the 2017 version of the training set, a combination of dev2010, tst2010, tst2015 as a validation set, and tst2017 as a test set. The resulting dataset consists of $232k$ ($50.3k$ with cross-sentential coreferences), $3.5k$, $1.2k$ examples of train, dev, test sets respectively.

[2] http://www.statmt.org/wmt19/translation-task.html

[3] https://wit3.fbk.eu/home

[4] https://opus.nlpl.eu/OpenSubtitles-v2018.php

OpenSubtitles We also choose the English-German pair of OpenSubtitles2018 corpora. The raw corpus contains 24.4M parallel sentences. We follow the filtering methods in [21] by removing pairs that have a time overlap of subtitle frames less than 0.9. We also use separate documents for validation / test sets, resulting in 3.9M (1.01M with cross-sentential coreferences), 40.7k, 40.5k examples for train / validation / test sets respectively.

En-Ko Subtitles For English-Korean experiments, we first crawled approximately 6.1k bilingual subtitle files from websites such as GomLab.com. Since sentence pairs of these subtitles are already soft-aligned by the creators so we applied a simple time-code based heuristics to filter examples. The final data contains 1.6M (0.24M with cross-sentential coreferences), 155.6k, and 18.1k examples of consecutive sentences in the training, validation, and test sets respectively.

For preprocessing, all English and German corpus is tokenized first with Moses [66] tokenizer[5]. We then apply the BPE [91] using SentencePiece[6], and the size of the merge operation is approximately 16.5k. We also put a special token [BOC] at the beginning of contextual sentences to differentiate them from the source sentences.

4.4.2 Settings

We use model hyperparameters, such as the size of hidden dimensions and the number of hidden layers as same the `transformer-base` [7], since all of the compared models are based on Transformer. Specifically, we set 512 as the hidden dimension, the number of layers is 6, the number of attention heads is 8, and the dropout rate is set to 0.1. On T5-based models (sent-level-t5, concat-t5), we used the `T5-Small` setting which has roughly the same number of parameters as `transformer-base` for fair comparison with others.

All models are trained with ADAM [65] with different learning rates for each

[5]https://github.com/moses-smt/mosesdecoder

[6]https://github.com/google/sentencepiece

dataset. We employ early stopping of the training when the MT loss on the validation set does not improve. We start training each baseline model from scratch with random initialization and document-level dataset. Note that all the baseline models are not trained using iterative training as [55, 72] which first trains the model from sentence-level task first, then document-level task. All the evaluated models including T5-based models are implemented on top of the transformers[7] framework.

We measure the translation quality by the BLEU score [28]. For scoring BLEU, we use the sacreBLEU [92] case-sensitive, detokenized scores for En-De, and case-insensitive scores with `intl` tokenizer for En-Ko task. We also report case-insensitive char-level scores on En-Ko for comparison.

4.4.3 Overall BLEU Evaluation

We display the corpus-level test BLEU scores of all compared models on different tasks in Table 4.1. Among the baseline systems, all context-aware models show moderate improvements over the sentence-level (sent-level) baseline. These results are comparable to that of [72] on the IWSLT task except for multi-enc-hier, and [24] on OpenSubtitles task. One exception is a single-encoder model (concat) on WMT task, which seems due to the longer average sentence length.

We evaluated CorefCL by fine-tuning the context-aware models. Results show that models with CorefCL outperformed their vanilla counterparts, with the BLEU gain of up to 1.4 in En-De tasks, and 1.6/2.8 (detokenized/char-level BLEU) in the En-Ko subtitles task. In addition to the models trained from scratch, CorefCL moderately improved the concat-t5 model on all the English-German tasks as well.

We observed that while CorefCL consistently improves BLEU on all tasks, it achieves better results on IWSLT and En-Ko subtitles tasks. Since improvements on much larger datasets like WMT and OpenSubtitles are smaller, we suggest that CorefCL also works as a regularization.

[7] https://github.com/huggingface/transformers

System	WMT	OpenSubtitles	IWSLT	En-Ko Subtitles	
				detok.	char.
sent-level	22.7	27.6	29.3	8.6	19.2
sent-level-t5	23.6	28.5	30.3	-	-
concat	22.4	28.3	29.7	9.3	22.1
+ CorefCL	23.5 (+1.1)	29.1 (+0.8)	30.9 (+1.3)	10.9 (+1.6)	24.9 (+2.8)
concat-t5	23.2	29.2	30.4	-	-
+ CorefCL	23.6 (+0.4)	29.5 (+0.3)	30.9 (+0.5)	-	-
multi-enc	23.1	28.6	29.8	9.2	21.7
+ CorefCL	24.3 (+1.2)	29.8 (+1.4)	31.1 (+1.3)	10.8 (+1.6)	24.4 (+2.7)
multi-enc-hier	24.4	29.1	30.0	10.3	23.1
+ CorefCL	25.4 (+1.0)	30.2 (+1.1)	31.1 (+1.2)	11.7 (+1.4)	25.7 (+2.6)

Table 4.1: Corpus-level BLEU scores of compared models on different tasks. For the En-Ko subtitles task, we list both detokenized (detok.) and character-level (char.) scores. Improvements by CorefCL are denoted in (). Underlined score means that the model has the largest BLEU improvements among models in the same task.

4.4.4 Results on English-German Contrastive Evaluation Set

System	Trained on			
	WMT		OpenSubtitles	
	BLEU	Acc.	BLEU	Acc.
sent-level	19.3	47.9	29.6	48.4
sent-level-t5	20.6	48.7	30.4	49.5
concat	19.9	49.7	30.5	54.4
+ CorefCL	20.3	51.2	32.3	57.9
concat-t5	20.9	50.2	30.9	56.3
+ CorefCL	21.2	52.0	32.5	58.7
multi-enc-hier	20.4	50.9	31.7	57.3
+ CorefCL	21.9	52.4	33.6	60.5

Table 4.2: BLEU and pronoun resolution accuracies on ContraPro [1] En-De contrastive test set.

To assess how CorefCL improves the ability to deal with pronoun-related translations more in detail, we experiment our method with ContraPro.[8] ContraPro is a contrastive test suit for En-De pronoun translation introduced by [1]. The evaluation is done by letting the model scores the German sentence with correct and incorrect pronoun translation, given the source and contextual English sentence. The accuracy is calculated by counting the number of correctly scored examples (i.e. correct examples that received a higher score than their incorrect counterpart).

We evaluate the models trained with WMT and OpenSubtitles tasks. We also list BLEU scores of En-De translation using the English source text in ContraPro. As shown in Table 4.2, CorefCL significantly improves the baselines in scoring accuracy for all models by up to 5.5%, as well as slight improvements in BLEU scores.

One interesting finding is that CorefCL also achieved substantial accuracy gain

[8] https://github.com/ZurichNLP/ContraPro

on the models trained on WMT. Since the ContraPro is created from OpenSubtitles, WMT-trained models would yield lower performance because of domain shift between training and testing. Table4.2 clearly shows the performance drop in BLEU, nevertheless, moderate improvements in accuracy can also be observed on WMT-trained models.

System	BLEU	Accuracy
multi-enc-hier	31.7	57.3
+ CorefCL	33.6	60.5
- Word omission	32.4	59.4
- Word replacement	32.3	58.6

Table 4.3: Ablation study on coreference corruption strategy. All systems are trained on OpenSubtitles English-German dataset and evaluated on ContraPro.

4.4.5 Analysis

Ablation Study CorefCL uses the two corruption strategies for generating contrastive coreference mentions; word omission and word replacement. To make a better understanding of influence of these strategies, we evaluate CorefCL of different settings of these strategies.

As shown in Table.4.3, using both types of corruptions results in better performance. Removing one of the two strategies slightly degrades both the pronoun resolution accuracy and BLEU. Although not being significant, removing the word replacement has more impact on accuracy. This suggests that a standard context-aware model, at least for multi-enc-hier is less sensitive to word substitution. The word replacement strategy can complement this behavior as resulted in better performance.

Qualitative Example We display a sample from ContraPro corpus and its translations made by multi-enc-hier model trained with OpenSubtitle task. In this example, since ”coat” is translated as *Mantel* which is a masculine noun thus *Er* would be ade-

Context	What'll I do with the coat? When you're through with it, send it to the police.
Source	It... It didn't belong to her.
multi-enc-hier	Sie... sie gehörte nicht zu ihr.
+ CorefCL	Er... er ist nicht ihr gehörte.
Reference	Er... er gehörte ihr nicht.

Figure 4.3: Example translation with and without CorefCL.

quate translation of ”It” instead of *Sie* which is feminine. While multi-enc-hier incorrectly translated ”It” as *Sie*, the model fine-tuned with CorefCL correctly resolved it as *Er*.

In practice, context-aware models that do not leverage target-side contexts struggle to maintain these kinds of coreference consistency [1, 93] because of the asymmetric nature of grammatical components and data distributions. Results show that CorefCL can complement the limitation of source-only context-aware models.

4.5 Summary of Contrastive Learning for Context-aware Neural Machine Translation

In this chapter, we have presented a data augmentation and contrastive learning scheme based on coreference for context-aware NMT. By leveraging coreference mentions between the source and target sentence, CorefCL effectively generates contrastive examples for applying contrastive learning on context-aware NMT models. In the experiments, CorefCL consistently improves the translation quality and pronoun resolution accuracy.

Chapter 5

Improving English-Korean Honorific Translation Using Contextual Information

Neural machine translation (NMT) has shown impressive results on translation quality, due to the availability of vast parallel corpus [94], and the introduction of novel deep neural network (DNN) architectures such as self-attentional networks [7]. The performance of NMT systems has reached on par with human translators in some domains, and hence many commercial MT services, such as Google Translation, have adopted NMT as their backbone of translation systems [8].

Despite the significant improvement over the previous machine translation (MT) systems, NMT still suffers from language-specific problems such as Russian pronoun resolution [17] and honorifics. Addressing such language-specific problems is crucial in both personal and business communications [95] not only because the preservation of meaning is necessary but also many of these language-specific problems are also closely related to their culture. *Honorifics* are good example of these language-specific problems that conveys respect to the audience. In some languages including Korean, Japanese, and Hindi that use honorifics frequently, speaking the right honorifics is considered imperative in those languages.

In Korean, one of the most frequent usages of honorifics occurs in the conversation

Sentence	English	Korean
context_1	Come on, **dad**. Don't you even take something?	**아빠**, 뭐라도 안 드세요?
context_0	Okay, give me some coffee.	좋아, 그럼 커피 좀 줘.
source/target	Wait a minute, please.	잠시만 기다려요.

Figure 5.1: An example of Korean dialogue that is extracted from subtitles. The blue words are verbs that translated into polite form whereas the red words are impolite form, using Korean honorifics.

with people who are in superior positions, or elders [2]. As is shown in Figure 5.1, the `source` English sentence "Wait a minute, please.", which is the second utterance by the son, is translated into the `target` sentence "잠시만 기다려요." (jam-si-man gi-da-lyeo-yo) that is represented as *haeyo-che* (해요체) as the sentence ends with -요 (-yo). Haeyo-che is a type of Korean honorific reflecting the relationship between the two speakers.

Addressing such honorifics in MT is challenging since the definition of honorifics differs across different languages. For example, Korean has 3 major types of honorifics [2] and corresponding honorific expressions. In contrast, it is known that English has fewer types of honorifics compared to many other languages [96]; only titles, such as Mr. and Mrs., are frequently used in modern English. It is known that managing honorifics in translation is comparatively more complicated in English-Korean translation; the source language has a simpler honorific system compared to the target language. The source language with fewer honorifics provides fewer honorific features that are used to generate correct honorifics in the target side, as shown in Figure 5.1. Since the English verb "wait" can be translated into both the honorific style (기다려요, gi-da-lyeo-yo) and the non-honorific style (기다려, gi-da-lyeo), the model cannot determine the adequate honorific solely depending on the source sentence, and additional information is necessary such as the relationship between speakers.

In this study, we propose a novel method to remedy limitations from solely depending on source sentence by using *context*, which is represented by the surrounding

sentences of the source sentence. In Figure 5.1, we can infer that this is a dialogue between a son and his father from the content of `context_1`, and the source sentence. Therefore, the model can determine that the source sentence should be translated into a polite sentence using honorifics, such as *haeyo-che* (해요체), if such context is taken into account.

To this end, we introduce a *context-aware* NMT to incorporate the context for improving Korean honorific translation. It is known that the context-aware NMT can improve the translation of words or phrases that need contextual information, such as pronouns that are sensitive to the plural and/or gender [97]. Considering above example that how the adequate honorific style can be determined using the context, we suggest that the context-aware NMT can also be used to aid the honorific-aware translation. To the best of our knowledge, this work is the first attempt to utilize context-aware NMT for honorific-aware translation.

We consider two types of context-aware NMT framework in our proposed method. First, we use a contextual encoder that takes context in addition to the source sentence as input. The encoder captures contextual information from the source language that is needed to determine target honorifics. Second, a context-aware post-editing (CAPE) system is adopted to take the context of translated target sentences for refining the sentence-level translations accordingly.

To demonstrate the performance of our method, an honorific-labeled parallel corpus is needed so we also developed simple and fast rule-based honorific annotation for labeling the test data. In the experiments, we compared our context-aware systems with context-agnostic models and we show that our method outperformed the context-agnostic baselines significantly in both the overall translation quality and translation of honorifics.

5.1 Related Works

5.1.1 Neural Machine Translation dealing with Korean

There have been a number of MT studies involving Korean. Because parallel corpora containing Korean are not as widely available as English and many European languages, a number of the existing works focused on low-resource MT settings. For example, Heo et al. [98] exploited out-of-domain and multilingual parallel corpora, and Jeong et al. [99] applied LM pretraining and back-translation. In addition, some other works have attempted to develop additional techniques to overcome the limitations of common low-resource MT methods. For example, Nguyen et al. [100] incorporated morphological information and word-sense disambiguation (WSD) on Korean source sentences to improve the translation into Vietnamese. Park et al. [101] focused on beam search decoding and experimented with various decoding settings including beam size to improve translation quality without re-training the target NMT model. Although low-resource MT methods are out of scope in this study, some methods including back-translation are closely related with our methods in training CAPE.

5.1.2 Controlling the Styles in NMT

Although the style of a generated translation also affects the quality of the machine translation, it has received little attention in the field of NMT. Since the source sentence contains insufficient information of the output style, most of the existing works have introduced a set of special tokens [102]. For example, to control the formality of the target sentence, one can add <F> at the beginning of the source sentence to translate formally or add <I> to translate informally. The model can attend to this token and extract the relevant linguistic features on training. This approach has been adopted in many subsequent works such as [103, 104]. Some other works have addressed this problem as domain adaptation that treats each style as a domain [105] or adopted multitask learning of the machine translation and the style transfer problem to address

the lack of a style-annotated parallel corpus [106], but the output is still controlled by the special tokens. By contrast, our approach can improve the honorific translation without using such kinds of special tokens by exploiting the contextual information of the surrounding text. In addition, our method can be combined with the methods using special tokens to further improve the accuracy of honorifics.

On the other hand, a few kinds of grammatical styles have addressed the style-controlled MT. The English formality [107] or the T-V distinction in European languages such as Spanish [95] are two common examples. Viswanathan et al. [95] have addressed the control of T-V distinction such as the use of a formal/informal form of second-person pronouns (*usted* vs. *tú*), as domain adaptation. Niu et al. [107] has shown that employing syntactic supervision can improve the control of English formality. Furthermore, few studies have addressed the honorifics of Asian languages such as Korean [104] and Japanese [108]. Wang et al. [104] used data labeling and reinforcement learning (RL) to enhance translation of Korean honorifics. However, they ignored contextual sentences and only relied on special tokens to control the honorifics.

5.1.3 Context-Aware NMT Framework and Application

The context-aware MT models focus on contextual information in the surrounding text [18] and either the context of the source or the target sentence can be considered. Exploiting the source side of contexts usually implements an additional encoder to represent the multiple contextual sentences efficiently [17, 22, 23]. On the other hand, the target-side contexts can be exploited by first translating a part of documents or discourses at the sentence level and then refining those translations. This can be implemented either by the use of multi-pass decoding or automatic post-editing (PE). The multi-pass decoder generates the translation at the sentence level first and then translates again by regarding the translated sentences as contexts [20, 21]. On the other hand, the context-aware PE corrects the common and frequent errors of sentence-level models by considering both the target sentence and its contexts [97]. In contrast to our

previous studies, we choose to use both sides of contexts as the target Korean context is helpful in correcting inconsistencies of Korean honorific translations.

On applications of context-aware MT, many of previous studies have been focused on improving pronoun resolutions such as choosing the correct gender or plural for pronouns. For example, Voita et al. [17, 97] have addressed the translation of Russian, and Müller et al. [1] are focused on German pronoun resolution. To the best of our knowledge, our work is the first attempt to use context-aware NMT to control grammatical styles such as honorifics.

5.2 Addressing Korean Honorifics in Context

In this section, we present an overview of the Korean honorifics system and how the contextual sentence can be used to infer appropriate honorifics for translation.

5.2.1 Overview of Korean Honorifics System

Asian languages such as Korean, Japanese, and Hindi are well-known as having rich honorific systems to express formality distinctions. Among those languages, the use of honorifics is extensive and also crucial in Korean culture. In practice, Korean speakers are forced to choose appropriate honorifics in every utterance, and failing to do that can induce serious social sanctions including school expulsion [2]. Moreover, it is known that Korean honorific systems are very sophisticated among the well-known languages thus teaching how to use Korean honorifics appropriately is also considered challenging in Korean as a Second Language (KSL) education [2, 109]

There are three types of Korean honorifics; subject honorification, object honorification, and addressee honorification.

Subject Honorification

In the subject honorification, the speaker honors the referent by using honorific suffixes such as '-시-'(-si-), case particles such as '-께서'(-kke-seo), and so on:

1. *철수*가 방에 들어가다. (cheol-su-ga bang-e deul-eo-ga-da; Cheolsoo goes to the room.)

2. [-15]*어머니*께서 방에 들어가신다. (eo-meo-ni-kke-seo bang-e deul-eo-ga-sin-da; My mother goes to the room.)

In contrast to (1), the speaker's 어머니 (eo-meo-ni, mother) in (2) is honored by the following case particle '께서' (-kke-seo) and the honorific suffix '-신-' (-sin-) at the verb 가다 (ga-da; *go*).

Object Honorification

Object honorification is often used when the referent of the object is of higher status (e.g., elder) to both the speaker and referent of the subject :

1. *철수*는 잘 모르는 것이 있으면 항상 **아버지**께 여쭌다. (*cheol-su*-neun jal mo-leu-neun geos-i iss-eu-myeon hang-sang **a-beo-ji**-kke yeo-jjun-da; Cheolsoo always ask his father about something that he doesn't know well.)

2. **아버지**는 휴대폰에 대해 잘 모르는 게 있으면 항상 *철수*에게 묻는다. (**a-beo-ji**-neun hyu-dae-pon-e dae-hae jal mo-leu-neun ge iss-eu-myeon hang-sang *cheol-su*-e-ge mud-neun-da; Cheolsoo's father always ask him about mobile phones that he doesn't know well.)

In the example (a), Cheolsoo's **아버지** (a-beo-ji, father) is in the superior position both to *철수* (Cheolsoo) and the speaker. Therefore, '여쭌다' (yeo-jjun-da), which is an honorific from of the verb *묻는다* (mud-neun-da, ask), is used.

Style and Name	Politeness	Formality	Example
합쇼체	High	High	날씨가 춥습니다.
(Hapsio-che; Deferential)			nal-ssi-ga chub-seub-ni-da
해요체	High	Low	날씨가 추워요.
(Haeyo-che; Polite)			nal-ssi-ga chu-wo-yo
하오체	Neutral	High	날씨가 춥소.
(Hao-che; Semiformal)			nal-ssi-ga chub-so
하게체	Neutral	Low	날씨가 춥네.
(Hagae-che; Familiar)			nal-ssi-ga chub-ne
반말체	Low	High	날씨가 추워.
(Banmal-che; Intimate)			nal-ssi-ga chu-wo
해라체	Low	Low	날씨가 춥다.
(Haela-che; Plain)			nal-ssi-ga chub-da

Table 5.1: Speech levels and sentence endings in Korean. Names are translated with respect to [2]. Each of the example sentences are a Korean translation of ”The weather is cold”. Each underllined sentence ending corresponds to their addressee honorific.

Adressee Honorification

Addressee honorifics are expressions of varying speech levels that are used to show politeness or closeness and are usually expressed as sentence endings in Table 5.1.

Despite that all 6 examples are translated as the same English sentence, each example has its own levels of formality and politeness and different usages. For example, ‘반말체’ (*banmal-che*) and ‘해라체’ (*haela-che*) are used between people with close relationships or used by the elderly when speaking to younger people. Conversely, ‘해요체’ (*haeyo-che*) and ‘합쇼체’ (*hapsio-che*) are used to honor the addressees and express politeness [2].

Sentence	English	Korean
context_1	You're back.	자네들 또 왔구만.
context_0	Yes, **sir**, we are.	예, **어르신**.
source/target	We're addicted to **your** citrus.	**어르신**의 감귤류에 중독 됐어요.

(a)

Sentence	English	Korean
context_1	You need to relax, okay?	진정해 주실래요?
context_0	You are not a **suspect**.	당신은 **용의자**가 아닙니다.
source/target	We should find Jessica right now.	저희는 빨리 제시카를 찾아야만 합니다.

(b)

Figure 5.2: Two examples of Korean dialogue from our dataset, which are extracted from subtitles. The blue words are verbs that translated with polite and/or formal honorifics whereas the red words are translated with impolite and/or informal honorifics. The bold keywords are used to determine what types of honorifics should be used. The underlined pronouns indicate that the two utterances is told by the same speaker in (**a**) and the utterances are formal speech in (**b**).

5.2.2 The Role of Context on Choosing Honorifics

As stated earlier, the relationship between speaker and audiences affects the use of Korean honorifics. For example, the student should use *haeyo-che* and *hapsio-che* as addressee honorifics when asking a teacher some questions. Since such social context is often reflected in utterances, readers may infer the relationship from text without knowing who are speakers and/or audiences.

In the Figure 5.1, we can infer that the source and the contextual sentence is consist of a dialogue between a dad and a son and the `context_1` and the source sentence is utterances of the son, so the source English sentence should be translated into a polite Korean sentence as shown.

Figure 5.2 shows two another examples in our dataset. In (a), a dialogue between a person (`context_0`) and his/her superior (`context_1`). So their Korean trans-

lations are in polite (*haeyo-che*) and impolite (*banmal-che*) respectively. In addition, we can infer that the source sentence is also an utterance by the same person who told (`context_0`) as we can find the same pronoun we to refer themselves. So the sentence endings of translation should be as "중독 됐어**요**" (jung-dog dwaess-eo-**yo**) which has the same honorifics as `context_0`, instead of using *banmal-che*, such as "중독 됐**어**" (jung-dog dwaess-**eo**).

On the other hand, (b) shows the usage of *hapsio-che* which is frequently used for formal expressions in `context_0` and the source sentence, as both of the sentences are ending with '-ㅂ니다' (-b-nida). The word *suspect* (용의자, yong-ui-ja) in `context_0` give us a hint that the `context_0` is told by police officers, prosecutors etc since the word is frequently used by those occupations. We can also infer that this dialogue is not held between those officers from the pronoun you, rather the utterances are told to a witness, etc. So the `context_0` and the source sentence would be translated into formal Korean utterances, rather than informal sentences like "우린 빨리 제시카를 찾아야 **해**" (u-lin ppal-li je-si-ka-leul chaj-a-ya **hae**).

As shown in the examples, contextual sentences often have important clues for choosing appropriate honorifics in Korean translation. However prior approaches for honorific-aware NMT including [108] for Japanese, and [104] for Korean have ignored those contexts. Instead, they explicitly controlled the model to translate the source sentence into a specific honorific style, using special tokens for indicating the target honorific as [102].

5.3 Context-Aware NMT Frameworks

To utilize the contextual sentences in NMT, we introduce the CNMT systems. These are divided into two categories: contextual encoders on NMT models and a CAPE system. Here we briefly review those systems before explaining our proposed method.

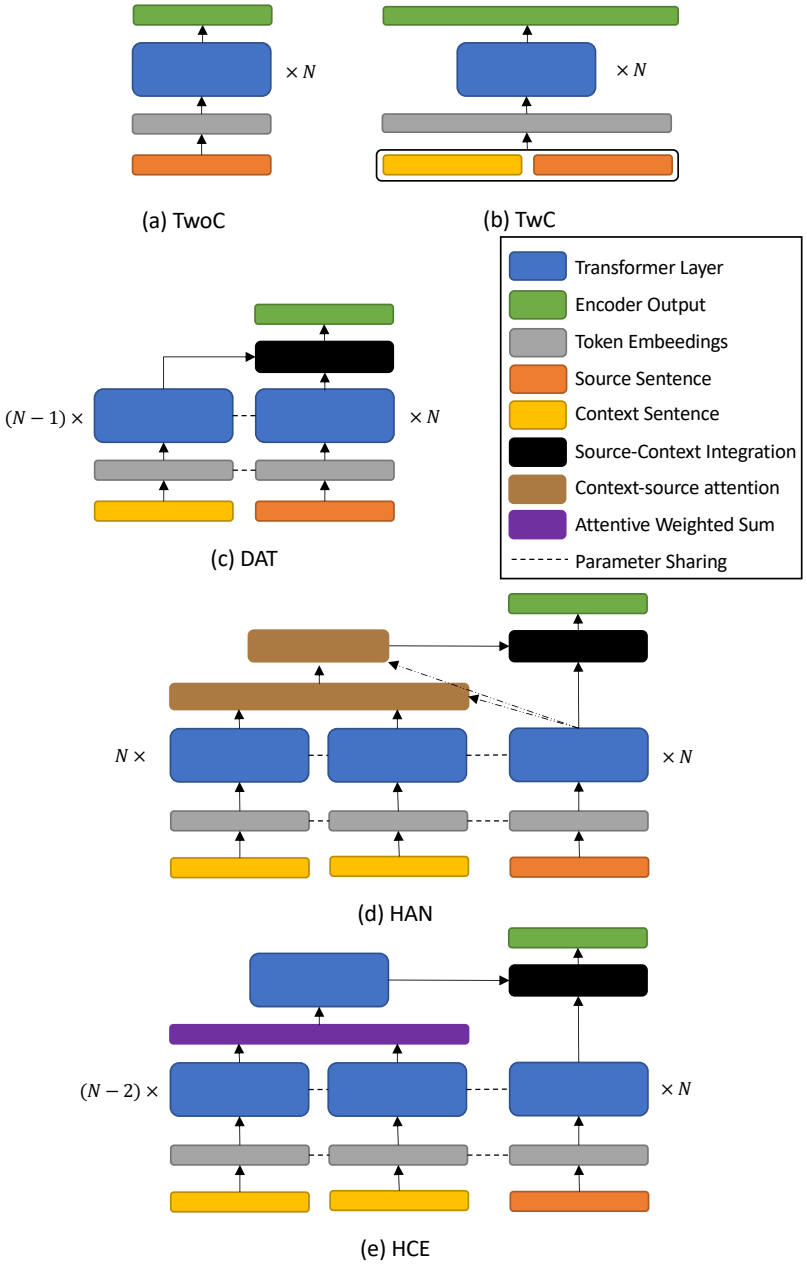

Figure 5.3: The structure of compared contextual encoders; (**a**) TwoC (**b**) TwC (**c**) DAT (**d**) HAN and (**e**) HCE.

5.3.1 NMT Model with Contextual Encoders

Generally, NMT models are operated at the sentence-level; it takes an input sentence in a source language and returns an output sentence in a target language. On the other hand, a contextual encoder in NMT is designed to handle one or more contextual sentences as input and extract the contextual representation. In our settings, NMT models are based on the self-attentional network (SAN) like Transformer [7]. Because of its strength in performance and efficiency, SAN has been widely used in NMT, and many improvements have also been made including contextual encoders. We list five SAN-based models in our experiments:

- **Transformer without contexts (TwoC)**: Vanilla sentence-level SAN as same as [7].
- **Transformer with contexts (TwC)**: SAN with a concatenation of the input and its contextual sentences as an input.
- **Discourse Aware Transformer (DAT)**: A multi-encoder model proposed by Voita el. al. [17].
- **Hierarchical Attention Networks (HAN)**: A multi-encoder model with hierarchical structure proposed by Miculicich et. al. [22].
- **Hierarchical Context Encoder (HCE)**: Our improved hierarchical multi-encoder model described in Chapter 3.

All the model structures are described in Figure 5.3. For detailed explanations on each model, please refer to the Section 3.2.1.

5.3.2 Context-Aware Post Editing (CAPE)

CAPE is a variant of automatic post-editing (PE) systems (e.g., Vu et al. [110]). The PE fixes systematic errors that frequently occur in a specific machine translation system.

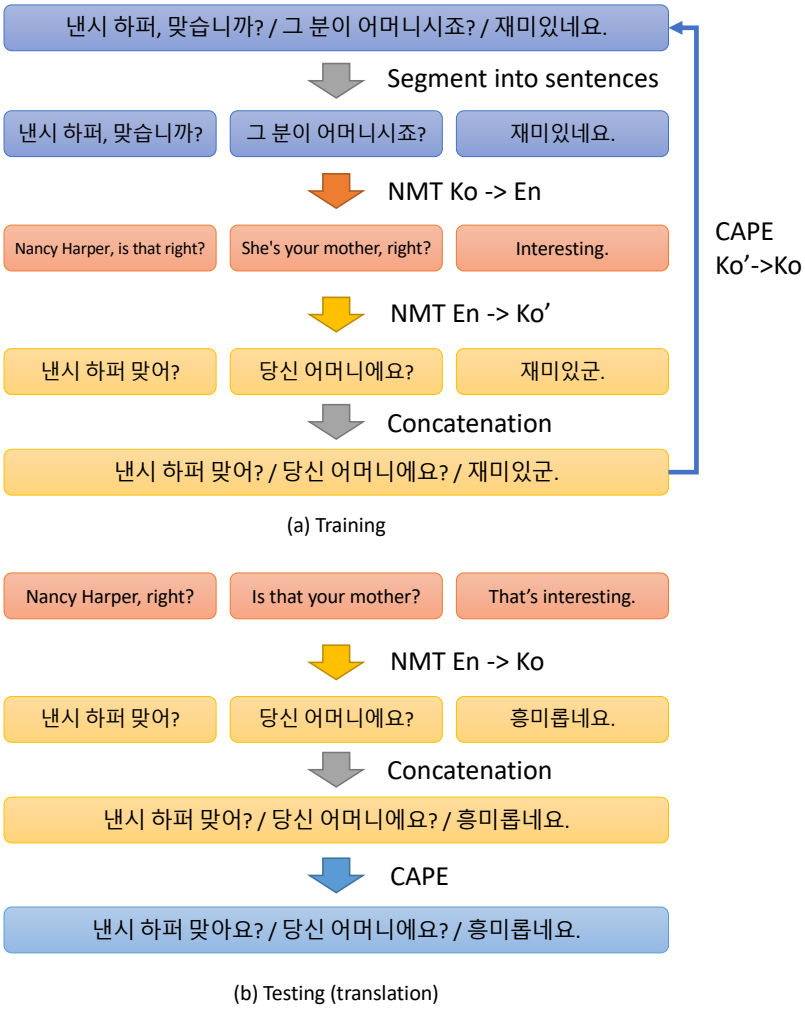

Figure 5.4: (**a**) Training a CAPE model requires a monolingual, discourse-/document-level corpus. Each consecutive text is segmented into a set of sentences first. Then, each sentence is translated and then back-translated. The resulting sentence group is concatenated again, and then the CAPE, which consists of a sequence-to-sequence model, is trained to minimize the errors of these round-trip translations. (**b**) At test time, a trained CAPE fixes sentence-level translations by taking them as a group.

Most of the PE operates at the sentence level; however, Voita et al. [97] suggested using PE to correct inconsistencies between sentence-level translations of a context-agnostic MT system. Analogous to many existing PE systems, the CAPE itself is independent of a specific MT model and can therefore in principle be trained to correct translations from any black-box MT system including a context-aware NMT system.

The training and testing process of CAPE is illustrated in Figure 5.4. First, the translation inconsistency of the target NMT model is simulated by using a round-trip translation. For example, to refine an English to Korean NMT system, Korean sentences are translated into English using Korean to English NMT first; then, they are again back-translated into Korean with a target English to Korean NMT system. In this way, the errors of the NMT model can be represented as the difference and inconsistency between the original Korean sentences and its round-trip translations. Once these round-trip translations are prepared, the CAPE, which consists of a typical sequence-to-sequence model, is trained to minimize these gaps. At test time, the target NMT system translates each sentence first, and then the CAPE takes a group of such translations and produces fixed translations. Moreover, CAPE has been shown to improve the English to Russian translation of context-sensitive pronouns [97] such as deixis and ellipsis.

5.4 Our Proposed Method - Context-Aware NMT for Korean Honorifics

In this section, we describe our proposed approach to generate appropriate Korean honorific expressions with context-aware NMT. We propose the use of context-aware NMT for translation of the honorific-styled sentence, which can improve the translation of honorifics without explicit control as done with special tokens. We also developed an automatic honorific labeling method to label the parallel corpus so that evaluation of the honorific translations, and preparing training data when the system is

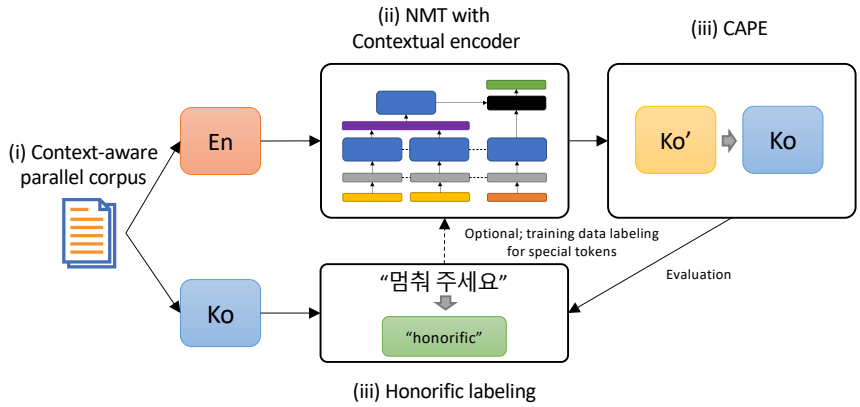

Figure 5.5: The process of our method, context-aware NMT for Korean honorifics. First we train NMT model with contextual encoder for English-Korean and Korean-English translation. Then we train CAPE to correct errors on those round-trip translations made by the NMT model. The automatic honorific labeling is primarily used for assessing honorific translation, but can also be used to label the training set if the NMT model uses special tokens to control target honorifics explicitly.

allowed to control target honorifics as in [104]. The process of our proposed method is illustrated in Figure 5.5.

5.4.1 Using CNMT methods for Honorific-Aware Translation

To capture contextual information that affects the use of Korean honorifics, our method exploits the context-aware models in two ways, as described in Section 5.3.

The first one is an NMT model with a contextual encoder (Section 5.3.1), which is trained to capture the dependency between the contents of contextual sentences of the source language and the usage of honorific expressions represented in the training data. For example, in Figure 5.1, the model can attend the noun *dad* in the `context_1` to generate a translation in *haeyo-che*. In this way, the trained model can implicitly control the translation to generate appropriate honorific expressions according to the

contextual sentences. In the experiments, we compare this approach against the NMT models that explicitly control the translation honorifics by introducing special tokens as in [104]. Furthermore, we adopted the CorefCL [27], a contrastive learning method proposed in Chapter 4 for boosting performance of these contextual encoder models. CorefCL makes the model more sensitive to content of the contextual sentences and it improved overall translation quality and En-De pronoun resolution.

The second one is a CAPE (Section 5.3.2) for improving the inconsistent sentence-level translation of honorifics. As stated earlier, the CAPE is trained by recovering inconsistent round-trip translations that require a pretrained bidirectional sentence-level MT model. Therefore, we first train a TwoC model to translate both Korean-English and English-Korean using the same parallel corpus. Then, we sample round-trip translations from a separately constructed monolingual Korean corpus and train a CAPE to reconstruct the original Korean sentence from the sampled round-trip translations, as illustrated in Figure 5.4. Our CAPE model is implemented using the same Transformer model as the TwoC [7], so once the monolingual corpus and its round-trip translations are prepared, training CAPE is similar to training a TwoC. We also apply the CAPE to improve the NMT models with contextual encoders, such as HCE. Despite that the CAPE was originally intended to correct the errors of sentence-level MT similar to TwoC [97], it can complement the NMT with a contextual encoder. Importantly, the CAPE exploits the context information of the target language, and some types of inconsistency, such as inter-sentence disagreement of honorifics, can only be identified in the target language. In the experiments, we show that the CAPE can further improve the honorific translation of HCE as well by correcting the inconsistency of honorifics between sentences.

5.4.2 Scope of Honorific Expressions

Our work focuses on the translation of addressee honorifics, which is a key factor in determining whether the sentence is *honorific style*. From the 6 types of sentence

(a) Original Sentence

타이머를 정지시킬 수 있겠어요?

(b) Morph/POS Tagging (c) Extract *eomi* (word ending)

('타이머', 'NNG'), ('를', 'JKO'), ('정지', 'NNG'),
('시키', 'XSV'), ('ㄹ', 'ETD'), ('수', 'NNB'),
('있', 'VV'), ('겠', 'EPT'), ('어요', 'EFN'), ('?', 'SF')

(d) Substring Matching

('어요', 'EFN')

Figure 5.6: Tagging sentences into honorific or nonhonorific styles. The original sentence (a) "타이머를 정지시킬 수 있겠어요?" (ta-i-meo-leul jeong-ji-si-kil su iss-gess-eo-yo; *Can you shut off the timer?*) is segmented into morphologies with their part-of-speech (POS) tags. Then we use 'eomi's to classify the sentence.

endings in Table 5.1, the *haeyo-che* and *hapsio-che* are usually considered honorific styles that are used frequently by age–rank subordinates speaking to superiors [2, 104]. Thus, we consider sentences having these two types of endings as *honorific sentences*, while others are *non-honorific* sentences. The target sentence in Figure 5.1 "잠시만 기다려**요**" (jam-si-man gi-da-lyeo-**yo**) whose ending is *haeyo-che*, is an example of an honorific sentence. In contrast, "잠시만 기다리**게**" (jam-si-man gi-da-li-**ge**) is a non-honorific sentence that is translated the same as in English according to our criteria since its ending is *hagae-che*.

5.4.3 Automatic Honorific Labeling

To assess the quality of honorific translation, we need to annotate the corpus into *honorific sentence* vs. *non-honorific* sentences. We developed heuristics using the above criteria to label the Korean sentences with honorific styles.

As illustrated in Figure 5.6, we first segment sentences into morphologies and obtain their part-of-speech (POS) tags. This ensures that our heuristic can correctly

identify the proper sentence ending. In our implementation, the Kkma Korean tagger [111] is used to extract morphologies and POS tags. Once morphologies and POS tags are extracted, we then select *eomi* (어미) which is the sentence ending. We picked morphologies whose tag starts with 'EF' [1] in our implementation. We label sentences as honorific if their *eomi* is *hapsio-che* or *haeyo-che*. In some cases where the morphology tagger fails to extract word endings, we resort to sub-string matching with sentence-ending markers such as '?', or '.' to correctly extract the proper sentence ending.

This heuristic is used primarily to label the test set for evaluation of our method; however, it can also be used to label the training set for training NMT models with explicit control of honorifics. In this case, the honorific label is used to generate a special token if the translation honorific of the model is controlled by a special token.

5.5 Experiments

To verify how the context-aware models improve Korean honorifics in English-Korean translation, we conduct comprehensive experiments and analyses on how context-aware MT models translate Korean honorifics. First, we constructed an English-Korean parallel corpus with contextual sentences. Then, we train and compare the models described in Section 5.3. Finally, a qualitative analysis is conducted on some examples from our proposed method.

5.5.1 Dataset and Preprocessing

To the best of our knowledge, there are no English-Korean discourse-level or context-aware parallel corpora that are publicly available. Thus, we constructed an English-Korean parallel corpus with contextual sentences. Basically we expand the English-Korean subtitles dataset introduced in Chapter 3 as these subtitles data contain many

[1] `http://kkma.snu.ac.kr/documents/index.jsp?doc=postag`

scripts with honorific expressions.

We first crawled subtitle files from websites such as GomLab.com. Combined with the data crawled for experiments in the Section 3.4, our raw subtitle set consist of approximately 6100 files. Then, we split these files into training, development, and test sets, which consist of $5.3k$, 500, and 50 files, respectively. We applied a file-based split to make sure that contextual sentences are only extracted from the same movie/episode. Unlike other datasets such as OpenSubtitles2018 [62], our subtitle files contain both English and Korean sentences, so extracting bilingual sentence pairs is straightforward; we used timestamp-based heuristics to obtain those pairs. The resulting sentence pairs are $3.0M$, $28.8k$, and $31.1k$ pairs for training, development, and test sets, respectively. Some of the raw samples from our test sets are shown in Table 5.7.

The contextual sentences are selected by using the timestamp of each subtitle, which contains the start time and end time in milliseconds. We assume that the sentences contain contextual information if they appear within a short period of time before the source sentence. Specifically, the start time of a contextual sentence is within K milliseconds from the start time of the source sentence. We set K as 3000 heuristically, and the maximum number of preceding contextual sentences is 2 for all experiments except those of Section 6.4.2. The final data contains $1.6M$, $155.6k$, and $18.1k$ examples of consecutive sentences in the training, development, and test sets, respectively.

For monolingual data to train the CAPE, we added $2.1M$ Korean sentences using an additional 4029 crawled monolingual subtitles. The resulting monolingual data consist of $5.1M$ sentences.

We finally tokenized the dataset using the wordpiece model [8], and the size of the vocabulary is approximately $16.5k$. We also put a special token `<BOC>` at the beginning of contextual sentences to differentiate them from the source sentences.

Start (ms)	End (ms)	English	Korean
		...	
646819	649786	Gives us a view inside the house.	집 안 모습을 보여줍니다.
649786	652102	We have a clear heat signature of the occupants.	안에 있는 사람의 열이 분명히 감지되고 있습니다.
652102	655129	Size, stance, says male. We think he's armed.	몸집, 자세를 보면 남자입니다. 무기를 소지한 것 같군요.
655129	656482	Any sign of a woman?	여자는 없나요?
656482	659841	Second body, right here, prone, giving off heat.	두 번째 사람, 바로 저기요, 누워 있는데, 열이 감지됩니다.
		...	

Figure 5.7: Example parallel sentence pairs extracted from bilingual subtitles.

5.5.2 Model Implementation and Training Details

For NMT models, we use model hyperparameters, such as the size of hidden dimensions and the number of hidden layers as the `transformer-base` [7], since all of the models in our experiment share the same Transformer structure. Specifically, we set 512 as the hidden dimension, the number of layers is 6, the number of attention heads is 8, and the dropout rate is set to 0.1. These hyperparameters are also applied to the CAPE model. For NMT models with additional encoders (DAT, HCE), we share the weights of encoders. All the evaluated models are implemented on top of the transformers[2] framework.

All models are trained with ADAM [65] with a learning rate of 1e-3, and we employ early stopping of the training when loss on the development set does not improve. We trained all of the models from scratch with random initialization, and we do not pretrain the model on a sentence-level task as in [22, 104].

5.5.3 Metrics

We measure the translation quality by BLEU scores [28]. For scoring BLEU, we use the sacreBLEU [92] with `intl` tokenizer for properly evaluating Korean. We first measure BLEU scores with original translations and we refer to these scores as *normal* BLEU scores. In addition, we also measure *tokenized* BLEU scores by tokenizing translations prior to scoring BLEU, as a common practice in the evaluation of Korean NMT [99].

For honorifics, we set the accuracy of honorifics as the ratio of translations with the same type of honorific style with respect to the reference translations. For example, if the reference translation of an English sentence "Yeonghee is cleaning." is "영희가 청소**해요**." (yeong-hui-ga cheong-so-**hae-yo**; *haeyo-che - honorific*) and the model translation is "영희가 청소**한다**." (yeong-hui-ga cheong-so-**han-da**; *banmal-che - non-honorific*), the translation is considered inaccurate.

[2] https://github.com/huggingface/transformers

5.5.4 Results

First, overall BLEU scores and honorific accuracy are compared among MT models with various types of contextual encoders. We also examine the varying performance of these models with respect to the number of contextual sentences and effects of CAPE for improving honorific translations.

Effect of Contextual Encoders

To evaluate the effect of contextual information on the translation of Korean honorifics, we first measure the performances of context-agnostic and context-aware models. The results are summarized on Table 5.2. As shown in the results, all the context-aware models (TwC, DAT, HAN, and HCE) outperform the context-agnostic model (TwoC) in terms of BLEU. The HCE shows a significant English-Korean BLEU improvement over TwoC of approximately 1.07/2.03 and the TwC, DAT, and HAN also show slight improvements. We later use Korean-English TwoC and HCE trained in this experiment for generating round-trip translations on CAPE experiment since the HCE performed best among the context-aware models in terms of BLEU.

We also experimented with the models on Korean-English BLEU using the same dataset for comparison. All the context-aware models again outperformed the context-agnostic model in this experiment. Note that BLEU scores are lower in all English-Korean experiments compared to Korean-English BLEU in the same dataset. This is mainly due to the morphological-rich nature of Korean and the domain of the dataset, which consists of spoken languages.

In addition to the BLEU scores, the context-aware models are also better in translation with correct Korean honorifics in English-Korean translation. In particular, the HCE has improved the honorific accuracy by 3.6%. Since showing politeness is considered important in Korean culture as discussed in Section 5.2.1, we also focus on the accuracy of the test sets which are polite target sentences. The TwC outperformed all other models in this set up to 4.81% compared to TwoC. The HAN and HCE

Models	BLEU		Accuracy	Accuracy
	En-Ko	Ko-En	All Test Set	Polite Targets
TwoC	9.16/12.45	23.81	64.34	39.27
TwC	9.6/13.2	24.35	66.85	**44.08**
DAT [17]	9.36/12.98	23.96	65.12	38.7
HAN [22]	9.50/13.08	24.54	66.3	42.26
HCE [24]	**10.23/14.75**	**26.63**	**67.94**	42.42

Table 5.2: English-Korean BLEU scores and accuracy (%) of honorifics for context-agnostic (TwoC) and context-aware (TwC, DAT, and HCE) NMT models. English-Korean BLEU scores are shown as (normal/tokenized) respectively. All the models are trained and tested without any honorific labels or explicit control of honorifics.

also showed significant improvement over TwoC, while the DAT's accuracy is slightly lower than that of TwoC. We believe that such differences derive from how the model utilizes contextual information. Since we only use the sequence-level cross-entropy (CE) as a training objective, the more compact representations of contextual encoders in DAT, HAN, and HCE can improve the main objective (translation quality), but considering the raw information of contextual sentences as in TwC could be more beneficial to honorific translation.

Models	BLEU	Accuracy	Accuracy
		All Test Set	Polite Targets
TwoC + Special Token	9.36/12.68	99.46	98.91
HCE + Special Token	**10.83/14.79**	**99.49**	**99.04**

Table 5.3: English-Korean BLEU scores (normal/tokenized) and accuracy (%) of honorifics for models with explicit control of honorifics by special tokens on the input. All the models are forced to obtain the translation with the honorific style of the reference sentence.

# Contextual Sents.	BLEU	Accuracy All Test Set	Accuracy Polite Targets
1	9.23/12.88	65.42	40.31
2	**10.23/14.75**	**67.94**	**42.42**
3	9.83/13.49	66.56	41.93
4	9.31/12.92	64.8	39.27
5	8.98/12.09	63.3	36.48

Table 5.4: English-Korean BLEU scores (normal/tokenized) and accuracy (%) by the number of contextual sentences on HCE

On the other hand, all of the results in Table 5.2 are from models that do not have any explicit control of honorifics and do not employ the honorific-annotated dataset. For comparison with prior works that forced the model to translate with specific honorifics as [104], we also include the results of NMT models with special tokens for controlling output honorifics in Table 5.3. In particular, the TwoC with special tokens is the same as the data labeling (DL) method in [104]. The training set was labeled the same as the test set, with the method described in Section 5.4.3. As shown in the results, both models are able to translate almost all the test examples with the same honorifics as their references, which is a similar result to that in [104]. Interestingly, both controlled models also improve the translation quality over their counterparts without control, and the HCE with special tokens again outperformed TwoC with special tokens on BLEU.

In summary, the context-aware NMT models can improve not only the translation quality but also the accuracy of honorifics. While their improvements are less significant compared to the honorific-controlled models, they can nevertheless exploit the contextual information to aid in the correct translation of honorifics.

Models	# Contextual Sents.	BLEU	Accuracy All Test Set	Accuracy Polite Targets
TwC	2	9.6/13.2	66.85	44.08
	5	8.23/11.41	61.21	38.05
DAT	2	9.36/12.98	65.12	38.7
	5	8.02/11.2	60.94	33.2
HAN	2	9.5/13.08	66.3	42.26
	5	8.55/11.74	63.1	36.6
HCE	2	10.23/14.75	67.94	42.42
	5	8.98/12.09	63.3	36.48

Table 5.5: English-Korean BLEU scores (normal/tokenized) and accuracy (%) by the number of contextual sentences on all of the context-aware NMT models

Effect of the Number of Contextual Sentences

The number of contextual sentences has a significant effect on the model performance since not all the contextual sentences are important in obtaining an adequate translation [112]. Such redundant information can hurt the performance. Since this number is dependent on the model and the data, we carry out experiments to examine the effect of the number of contextual sentences. As shown in Table 5.4, both the BLEU and accuracy of honorifics are the best on 2 contextual sentences, and then they decay as the number increases. Similar effects are also shown by the other context-aware NMT models, as displayed in Table 5.5.

Effect of CAPE

We measure the effect of CAPE and results are provided in Table 5.6. The CAPE improved TwoC by 0.87/1.93 on BLEU and outperformed TwC and DAT on honorific accuracies by approximately 3 to 4%. The improvement in honorific accuracy suggests that CAPE can also repair the inconsistency of honorifics. We additionally applied

Models	BLEU	Accuracy All Test Set	Accuracy Polite Targets
TwoC	9.16/12.45	64.34	39.27
+CAPE	**10.03/14.38**	**67.5**	**43.81**
HCE	10.23/14.65	67.94	42.42
+CAPE	**10.55/15.03**	**69.16**	**46.51**

Table 5.6: English-Korean BLEU scores (normal/tokenized) and accuracy (%) of honorifics for models with/without CAPE.

CAPE to HCE. The result shows that HCE with CAPE also outperformed the vanilla HCE, supporting our hypothesis.

Effect of Contrastive Learning

Models	BLEU	Accuracy All Test Set	Accuracy Polite Targets
HCE	10.23/14.65	67.94	42.42
+CL	11.64/16.49	68.72	45.3
+CAPE	10.55/15.03	69.16	46.51
+CL+CAPE	**12.1/17.27**	**71.34**	**47.26**

Table 5.7: English-Korean BLEU scores (normal/tokenized) and accuracy (%) of honorifics for models with CorefCL (denoted as +CL), and CAPE (as +CAPE). Using both the CorefCL and CAPE (as +CL+CAPE) results in the best performance.

Finally, we investigate how the contrastive learning (CorefCL) introduced in Chapter 4 can improve the overall BLEU and honorific accuracy on HCE. The results displayed in Table 5.7 clearly show improvements made by CorefCL either with or without CAPE, and using both the CorefCL and CAPE recorded the best performance. One interest result is that HCE+CL excels in BLEU compared to HCE+CAPE, but

with lower honorific accuracies. We suggest that this may due to a data augmentation strategy of CorefCL that does not exploit honorifics.

5.5.5 Translation Examples and Analysis

We show some translation examples in Figures 5.8 and 5.9. As discussed in Section 5.4, the honorific sentences are mostly used when a subordinate such as a child is talking to superiors such as his/her parents. Figure 5.8 shows two examples of these situations. In (a), context and source sentences are a conversation between a mother and her child. This can be speculated from the contextual sentences; the child is talking but the mom urges him/her to continue eating. The TwoC completely ignores the contextual sentences, so such a situation is not considered. Thus, TwoC translates the source sentence as a non-honorific style using the non-honorific sentence ending 때 (ttae), which is *banmal-che*. In contrast, the translation of HCE is an honorific sentence since its sentence ending is 요 (yo), which is *haeyo-che*, the same as the reference. This is an example that shows HCE's context-awareness that helps translation of honorific-styled sentences.

On the other hand, *Daddy!* in `context_1` of (b) and the content of `context_1` directly indicate that the source sentence is spoken by a dad's child. Despite such direct hints, HCE failed to correctly identify the proper honorific style, resulting in *banmal-che* (*해* (hae) and *어* (eo)). However, the TwC correctly translated the source sentence as an honorific sentence using *haeyo-che* (*해요* (haeyo) and *데요* (daeyo)). Note that there are two sentence segments in the source and translations, and the honorific style of the two segments agrees in all the model translations and the reference. One interesting observation is that TwC has translated verb sorry as 죄송-하다 (joesong-hada) instead of 미안-하다 (mian-hada) and the 2nd person pronoun **you** as **아빠/(appa; daddy)** instead of 네 (ne; you) like HCE. As the former is resulting as a more polite translation and the latter is closer to the reference so this example can be viewed as a clue that TwC's context-awareness is better than that of HCE. We suggest that TwC's

En (Context_1)	Life must go on as it always has. (언제나처럼 인생은 계속되어야죠.)
En (Context_0)	Come on, let's eat. (어서 먹자.)
En (Source)	How's mom?
Ko (TwoC)	엄마는 어때?
Ko (HCE)	엄마는 어떠세요?
Ko (Reference)	엄마는 어때요?

(a)

En (Context_1)	Daddy! (아빠!)
En (Context_0)	Hey, I'm here. (어, 나 여기 있어.)
En (Source)	I'm sorry. I should have listened to **you**.
Ko (TwoC)	미안해. **네** 말을 들었어야 했는데.
Ko (HCE)	미안해. **네** 말을 들었어야 했어.
Ko (TwC)	죄송해요. **아빠** 말을 들었어야 했는데요.
Ko (Reference)	미안해요. **아빠** 말을 들을 걸 그랬어요.

(b)

Figure 5.8: Example translations of different NMT models. The sentences are given in a sequence, from context_1 to source. The reference translation of each contextual sentence is given in (). In (**a**), a mother and her child are talking to each other. The context-aware model (HCE) can infer this situation using contextual sentences and translate the source sentence with an appropriate honorific style. Similarly, in (**b**) a dad and his child are talking, but only a translation from TwC has the correct honorific style. Note that translations of the verb sorry and the 2nd person pronoun **you** also differ among models despite that all the translations have the same meaning as the source sentence.

En	My condolences. / Skip the sympathy. This is a business. / My father met with you right before he died.
Ko (TwoC)	조의를 표합니다. / 조문은 필요없어. 이건 사업이야. / 아버지가 죽기 직전에 널 만났어.
Ko (HCE)	명복을 빕니다. / 조문은 필요없어. 이건 사업이야. / 아버지가 죽기 직전에 당신을 만났어요.
Ko (HCE+CAPE)	명복을 빕니다. / 조문은 필요없어요. 이건 사업이에요. / 아버지께서 돌아가시기 직전에 당신을 만났어요.
Ko (Reference)	고인의 명복을 빕니다. / 조문은 필요없어요. 이건 비즈니스입니다. / 아버지께서 돌아가시기 직전에 당신을 만났죠.

Figure 5.9: Example of a translation made by HCE and its correction by CAPE. The second and third sentence segments are the utterance of the same speaker. HCE's translations are inconsistent in honorifics since honorifics of the second and third segments do not agree. The CAPE successfully corrected that inconsistency. Note that CAPE also fixed the subject honorification, resulting in a more polite translation. Note that the underlined nouns are differ among models, despite that all the translations have the same meaning.

simple and direct use of contextual sentences can perform better than the abstract representation of contextual sentences in HCE when the contextual sentences are simple and short.

Finally, Figure 5.9 shows how the CAPE corrects the inconsistent use of honorifics. These 3 sentence segments are obtained from a scene held in a funeral home. Considering the content of the sentences, we can assume that the 2nd and 3rd segments are the utterances of the same speaker. However the honorific styles of HCE translations do not agree on *banmal-che* for the 2nd segment and *haeyo-che* for the 3rd. CAPE corrected this inconsistency by looking at the translated Korean sentences. In addition, CAPE also amended the 3rd sentence segment by modifying the subject honorification, replacing both the case particle for the subject (his father) from *-가* (-ga) to *께서* (-kkeseo) and the verb *죽기* (jukgi) to *돌아가시기* (doragasigi); both are translated as *died*. Considering that a deceased person is generally highly honored in Korean culture, the CAPE's correction results in a more polite and thus adequate honorific-styled sentence. Although the subject honorification is out of scope in this study, this shows the CAPE's ability to capture various honorific patterns observed in the training corpus and correct translations.

5.6 Summary of Improving English-Korean Honorific Translation Using Contextual Information

In this chapter, we have introduced the use of context-aware NMT to improve the translation of Korean honorifics. By using contextual encoders, the context-aware NMT models can implicitly capture the speaker information and translate the source sentence with proper honorific style. In addition, context-aware postediting (CAPE) is adopted to improve the honorific translation by correcting the inconsistent use of honorifics between sentence segments. Experimental results show that our proposed method can improve the translation of Korean honorifics compared to context-agnostic methods

both in BLEU and honorific accuracy. We also demonstrated that the use of context-aware NMT can further improve the prior methods which use special tokens to control honorifics translation. Qualitative analysis on sample translations supports the effectiveness of our method on exploiting contextual information for improving translations of honorific sentences.

Chapter 6

Future Directions

While many advances have been made in document-level NMT and CNMT recently, there are still many issues for exploration. Challenges on document-level NMT are not limited to better modeling of the document-level context but include developing better document-level datasets, context-dependent evaluations, and addressing ethical and practical issues. Here we discuss a few of the possible future research directions.

6.1 Document-level Datasets

Since most of the publicly available MT datasets are aligned in sentence-level, the need on creating document-level bilingual datasets is increasing. There have been several efforts to create new document-aligned datasets [113] or extending existing corpus by adding document boundaries [90].

However, there are still a number of issues that exist in those curations. One of the problems is that many of the current public datasets, especially for training sets do not have annotations on context-dependent discourse phenomena. Expanding automatic annotation of discourse phenomena such as coreference as introduced in Chapter 4 would be a viable direction as this explicit information could help improve lexical cohesion.

In addition, domains and language pairs of datasets should also be extended into multi-domain and morphologically rich languages, which are more challenging settings and would reflect practical applications of document-level NMT methods. It is the right time to invest efforts in creating such resources so that researches can be standardized with respect to the datasets used.

6.2 Document-level Evaluation

Sentence-level metrics like BLEU and METEOR are still used to evaluate MT systems, as they are widely accepted in the MT community for almost two decades. Since these metrics have limitations in assessing context-dependent discourse phenomena, developing and improving evaluation methods is crucial for making progress in document-level MT. Despite a number of document-level metrics and discourse phenomena test suits have been proposed, currently, there is no consensus about the evaluation of document-level MT. This is because many of the proposed document-level metrics including [115, 114] have their own limitations, and test suits like [1] only cover a part of the problem since they are mostly designed for specific language pairs.

For evaluation metrics, pre-trained language model(PLM)-based metrics like BERTScore [116] have attracted some attention recently. These metrics have advantages in taking the linguistic capabilities of PLMs and addresses several flaws of metrics like BLEU by not relying solely on n-grams. Currently, these PLM-based metrics are mostly work in the sentence-level, so extending the method to document-level would be an interesting option. In addition, an alternative would be found between automatic and manual evaluation that could enable a more economical manual evaluation, and it would still be better than the current automatic metrics at assessing discourse phenomena.

On the other hand, more attention are needed on extending test suits to cover more language pair and discourse phenomena, especially for non-European languages. Most

of the recent efforts have been focused on pronoun-related problems and language pairs involving English and European languages.

6.3 Bias and Fairness of Document-level NMT

Recently the problem of biased models has been recognized as important in the field of machine learning [117] since it hinders the fairness of the model. Such biases like gender bias are frequently induced from training data and NMT is also affected [118].

For example, when translating "*The engineer asked for diagrams.*" to a language with grammatical gender like Spanish, it is necessary to determine the gender of the subject "*engineer*" for obtaining adequate translation. If the sentence does not contain gender information, machine translation systems generally select the most common translation option learned from the data. These translations often correspond to the stereotypical translations as pointed in [119] that the MT service frequently mistranslated the female author into masculine pronouns. Such biases potentially induce several harms including exacerbation of prejudice.

As discussed in Savoldi et al.'s survey [120], several studies have addressed mitigating gender biases in NMT, such as adding extra information to preserving gender [118] or introducing debiasing techniques such as balanced fine-tunings [121]. Among these approaches, CNMT has also been used to incorporate gender information that can be inferred from contextual sentences. In Basta et al.'s study [122], the CNMT model with 1 preceding sentence significantly improved gender accuracies on English-Spanish WinoMT test suit [123] over sentence-level baselines. This improvement is quite impressive as the gender disambiguation addressed in WinoMT is more challenging than simply maintaining grammatical gender like the example in Figure 1.1.

Despite the promising results, applications of CNMT for reducing gender bias need more exploration as studies on the identification and evaluation of gender bias are still

emerging. For example, WinoMT mainly focuses on occupation and pronoun-related biases and there are recent approaches to introduce extended test suits to cover broader categories of biases such as stereotypical adjectives [124]. In addition, improvements by CNMT should be carefully examined since there are significant bias-unrelated improvements by CNMT as pointed in [122].

6.4 Towards Practical Applications

There are a number of issues facing real-world applications of NMTs. Many of these research topics have not yet been widely addressed in the venue of document-level NMT. The following topics are non-exhaustive, yet to be considered as guidelines for future researchers.

Domain Adaptation Generally NMT models perform poorly on out-of-domain data. For example, a model trained on news data is unlikely to achieve good performance on the biomedical corpus. As it is impractical to train the model on all domains of data and there is always the possibility of additional domains at a later stage (e.g. translation of newer texts with older model), re-purposing the model, or domain-adaptation for newer data is desirable. Domain adaptation is can be implemented in the training and/or inference process [125]. A few works have been applied domain adaptation on CNMT settings [126] for improving multi-domain performance.

Automatic Post-editing As stated in Chapter 5, automatic post-editing (APE) aims to improve the quality of an existing, black-box MT system by using human-edited examples. Recent researches have focused on neural APE systems leveraging transfer learning and data augmentation [127]. Although APE has been proposed in the CNMT framework as [97], this does not leverage human-edited data and relied on a simple sequence-to-sequence model and training scheme.

Translation Using Terminologies Domains like biomedical require very careful use of terminologies to get adequate translations. These domains also often face data

scarcity issues, especially on lack of high-quality parallel corpus enough to train NMT models. One good example is the recent COVID-19 pandemic which has forced urgent translation of the latest medical information. Instead of relying solely on parallel bitexts, exploiting word- or phrase-level dictionaries of key terms is an emerging area of research [128] as this setting is also prevalent and prominent on commercial MT services. Many of the existing approaches involve pre- and post-processing of sentences since incorporating dictionaries directly into the NMT model is nontrivial. On the other hand, document-level MT could be aided by these schemes since maintaining coherence on terminology and entity is an important aspect.

Chapter 7

Conclusions

This dissertation mainly focused on improving document-level translation quality of the context-aware neural machine translation (CNMT) model and investigating its application to deal with language-specific problem. To that end, we first tackle the efficiency issues on modeling multiple contextual sentences on encoder. We introduce a hierarchical context encoder (HCE) in Chapter 3 that encodes multiple contextual sentences from token-level to sentence-level. By adapting the hierarchical structure, the HCE consuming less computation time on training and translation than existing state-of-the-art CNMT encoders.

Secondly, we investigate training process for CNMT models, where most models rely on negative log-likelihood (NLL) that do not fully exploit contextual dependencies. To overcome the insufficiency, we introduce a coreference-based contrastive learning for CNMT in Chapter 4 that generates contrastive examples from coreference chains between the source and target sentences. This novel training method enables the model to generate more adequate pronoun translation, as well as improve overall translation quality.

Finally, we move on an application of CNMT, that is dealing with Korean honorifics translation which requires contextual information for adequate translations in Chapter 5. Since this problem needs the model to exploit contexts of both the source

and the target languages, we propose to use CNMT models that captures crucial contextual information on English source document and thee conext-aware post-editing (CAPE) system for exploiting contexts on Korean target sentences. We also design a Korean honorific test suit for assessing the models's ability on translating adequate Korean honorifics. Empirical results shows our method's strength in more consistent Korean honorific translations.

To sum up, this dissertation presents better approaches on CNMT model architecture and training method. In addition, we provide an application of CNMT that dealing with lanugage-specific problem and future research directions on document-level MT. We hope it improves the overall quality of document-level machine translation and thus expanding the real-world use of CNMT, especially in communicating with Korean. We suggest that spoken language translation (SLT) systems for applications like video captioning, online conference, and messaging can be benefited from our method.

Bibliography

[1] M. Müller, A. Rios, E. Voita and R. Sennrich, "A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium, 2018. pp. 61–72.

[2] L. Brown, "Questions of appropriateness and authenticity in the representation of Korean honorifics in textbooks for second language learners," *Language, Culture and Curriculum*, vol. 23, no. 1, pp. 35-50, 2010.

[3] P. Brown, J. Cocke, S.D. Pietra, V.D. Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin., "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no.2, pp. 79-85, 1990.

[4] P. Brown, S. Della Pietra, V. Della Pietra and R. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no.2, pp. 263-311, 2003.

[5] K. Cho, B. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, October 2014. pp. 1724-1734.

[6] I. Sutskever, O. Vinyals and Q.V. Le, "Sequence to Sequence Learning with Neural Networks" in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montréal, Canada, December 2014. pp. 3104-3112.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, USA, December 2017. pp. 6000-6010.

[8] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao and K. Macherey et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," arXiv preprint arXiv:1609.08144, September 2016.

[9] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser, "Universal transformers," in *Proceedings of the 6th International Conference on Learning Representations*, New Orleans, USA, April 2019.

[10] D. Jurafsky and J.H. Martin, *Speech and Language Processing (2nd ed.)*, Prentice-Hall, Inc., Upper Saddle River, 2009.

[11] R. Sennrich, "Why the time is ripe for discourse in machine translation?," in *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, Melbourne, Australia, July 2018.

[12] S. Maruf, F. Saleh, and G. Haffari, "A Survey on Document-level Neural Machine Translation: Methods and Evaluation," *ACM Computing Surveys*, Vol. 54, No. 2, Article 45, March 2021.

[13] K.S. Smith, "On Integrating Discourse in Machine Translation," in *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark, September 2017. pp. 110–121.

[14] S. Jean, S. Lauly, O. Firat, and K. Cho, “Does Neural Machine Translation Benefit from Larger Context?,” arXiv preprint arXiv:1704.05135, April 2017.

[15] S. Kuang and D. Xiong, “Fusing recency into neural machine translation with an inter-sentence gate model,” in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, USA, August 2018. pp. 607–617.

[16] L. Wang, Z. Tu, A. Way, and Q. Liu, “Exploiting cross-sentence context for neural machine translation,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* , Copenhagen, Denmark, September 2017. pp. 2826–2831.

[17] E. Voita, P. Serdyukov, R. Sennrich, and I. Titov, “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)*, Melbourne, Australia, July 2018. pp. 1264–1274

[18] J. Tiedemann and Y. Scherrer, “Neural Machine Translation with Extended Context”. in *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark, September 2017. pp 82–92.

[19] R. Bawden, R. Sennrich, A. Birch, and B. Haddow. 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*, New Orleans, USA, June 2018. pp. 1304–1313.

[20] H. Xiong, Z. He, H. Wu, and H. Wang, “Modeling Coherence for Discourse Neural Machine Translation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, USA, January 2019. pp. 7338-7345.

[21] E. Voita, R. Sennrich, and I. Titov, “When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lex-

ical Cohesion," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019. pp. 1198-1212.

[22] L. Miculicich, D. Ram, N. Pappas, and J. Henderson, "Document-Level Neural Machine Translation with Hierarchical Attention Networks," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018. pp. 2947–2954.

[23] S. Maruf, A.F.T. Martins, and G. Haffari, " Selective Attention for Context-aware Neural Machine Translation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, USA, June 2019. pp. 3092-3102.

[24] H. Yun, Y. Hwang, and K. Jung, "Improving context-aware neural machine translation using self-attentive sentence embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, New York City, USA, February 2020. pp. 9498–9506.

[25] Y. Kim, D.T. Tran, and H. Ney. "When and why is document-level context useful in neural machine translation?," in *Proceedings of the Fourth Workshop on Discourse in Machine Translation*, Hong Kong, China, November 2019. pp. 24–34.

[26] B. Li, H. Liu, Z. Wang, Y. Jiang, T. Xiao, J. Zhu, T. Liu, and C. Li, "Does multi-encoder help? a case study on context-aware neural machine translation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. pp. 3512–3518.

[27] Y. Hwang, H. Yun, and K. Jung. "Contrastive Learning for Context-aware Neural Machine Translation Using Coreference Information," in *Proceedings of the Sixth Conference on Machine Translation*, Punta Cana, Dominican Republic, November 2021.

[28] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Philadelphia, USA, July 2002. pp. 311–318.

[29] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, USA, June 2006. pp. 65–72.

[30] Y. Hwang, Y. Kim and K. Jung, “Context-Aware Neural Machine Translation for Korean Honorific Expressions,” *MDPI Electronics* vol. 13, no.1, May 2021.

[31] P. Koehn, F.J. Och and D. Marcu, “Statistical Phrase-Based Translation,’’ in *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada, May 2003. pp. 127-133.

[32] A. Castaño, F. Casacuberta and E. Vidal, “Machine Translation using Neural Networks and Finite-State Models,’’ in *Proceedings of the 7th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Santa Fe, USA, July 1997.

[33] Y. Goldberg, “A Primer on Neural Network Models for Natural Language Processing,’’ *Journal of Artificial Intelligence Research*, vol. 57, no. 1, pp. 345–420, 2016.

[34] H. Schwenk, “Continuous space language models,’’ *Computer Speech and Language*, vol. 21, no. 3, pp. 492–518, 2007.

[35] P. Li, Y. Liu, M. Sun, T. Izuha and D. Zhang, “A Neural Reordering Model for Phrase-based Translation,’’ in *Proceedings of COLING 2014, the 25th Interna-*

tional Conference on Computational Linguistics: Technical Papers, Dublin, Ireland, August 2014. pp. 1897-1907.

[36] H. Schwenk, “Continuous Space Translation Models for Phrase-Based Statistical Machine Translation,’’ in *Proceedings of COLING 2012: Posters*, Mumbai, India, December 2012. pp. 1071–1080.

[37] R. Sennrich, B. Haddow, and A. Birch. “Improving neural machine translation models with monolingual data,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, August 2016. pp. 86–96.

[38] S. Jean, O. Firat, K. Cho, R. Memisevic and Y. Bengio, “Montreal Neural Machine Translation Systems for WMT’15,’’ in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September 2015. pp. 134–140.

[39] O. Bojar et al., “Findings of the 2017 Conference on Machine Translation (WMT17)’’, in *Proceedings of the Second Conference on Machine Translation,* Copenhagen, Denmark, September 2017. pp. 169-214.

[40] P. Koehn, “Neural Machine Translation,’’ arXiv preprint arXiv:1709.07809, September 2017.

[41] J. Gu, J. Bradbury, C. Xiong, V.O.K. Li and R. Socher, “Non-Autoregressive Neural Machine Translation,” in *Proceedings of 6th International Conference on Learning Representations*, Vancouver, Canada, April 2018.

[42] J. Gu, J. Bradbury, C. Xiong, V.O.K. Li and R. Socher, “Fully Non-autoregressive Neural Machine Translation: Tricks of the Trade,” arXiv preprint arXiv:2012.15833, December 2020.

[43] J. Song, S. Kim, S. Yoon, “AligNART: Non-autoregressive Neural Machine Translation by Jointly Learning to Estimate Alignment and Translate” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Punta Cana, Dominican Republic, November 2021. pp. 1-14.

[44] D. Bahdanau, K. Cho and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in *Proceedings of 3rd International Conference on Learning Representations*, San Diego, USA, May 2015.

[45] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory” *Neural Computation,* vol.9, no. 8, pp. 1735-1780, 1997.

[46] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y.N. Dauphin. “Convolutional sequence to sequence learning,” in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, July 2017. pp. 1243–1252.

[47] M.T. Luong, H. Pham, and C. Manning, “Effective approaches to attention-based neural machine translation,” in *Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, September 2015. pp. 1412-1421.

[48] B. Zoph, D. Yuret, J. May and K. Knight, “Transfer Learning for Low-Resource Neural Machine Translation,’’ in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, November 2016. pp 1568-1575.

[49] J. Devlin, M-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, USA, June 2019. pp. 4171–4186.

[50] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, "Language Models are Few-Shot Learners,'' arXiv preprint arXiv:2005.14165, July 2020.

[51] K. Song, X. Tan, T. Qin, J. Lu and T-Y. Liu, "MASS: Masked Sequence to Sequence Pre-training for Language Generation," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, USA, June 2019. pp. 5926-5936.

[52] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," arXiv preprint arXiv:1910.13461, Oct 2019.

[53] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P.J. Liu, "Exploring the Limits of Transfer Learning With a Unified Text-to-text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp 1–67, 2020.

[54] A. Graves, "Sequence Transduction with Recurrent Neural Networks," in *Proceedings of the Representation Learning Workshop at the International Conference on Machine Learning*, Edinburgh, Scotland, July 2012.

[55] J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, M. Zhang, and Y. Liu, "Improving the Transformer Translation Model with Document-level Context," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018. pp. 533–542.

[56] G. Tang, M. Muller, A. Rios, and R. Sennrich, “Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018. pp. 4263–4272.

[57] K. Tran, A. Bisazza, and C. Monz, “The Importance of Being Recurrent for Modeling Hierarchical Structure,” arXiv preprint arXiv:1803.03585, March 2018.

[58] S. Maruf, and G. Haffari, “Document Context Neural Machine Translation with Memory Networks,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018. pp. 1275–1284.

[59] Z. Tu, Y. Liu, Z. Lu, X. Liu, and H. Li, “Context gates for neural machine translation,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 87–99, 2017.

[60] Z. Lin, M. Feng, C.N.D. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, “A structured self-attentive sentence embedding,” arXiv preprint arXiv:1703.03130, March 2017.

[61] M. Cettolo, M. Federico, L. Bentivogli, J. Niehues, S. Stuker, K. Sudoh, K. Yoshino, and C. Federmann, “Overview of the IWSLT 2017 Evaluation Campaign,” in *Proceedings of the 14th International Workshop on Spoken Language Translation*, Tokyo, Japan, Dec 2017. pp. 1–14.

[62] P. Lison, J. Tiedemann and M. Kouylekov, “Opensubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan, May 2018.

[63] M. Schuster and K. Nakajima, “Japanese and Korean voice search,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, volume 1*, Kyoto, Japan, March 2012. pp. 5149–5152.

[64] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. Gomez, S. Gouws, L. Jones, L. Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer and J. Uszkoreit, “Tensor2tensor for Neural Machine Translation,” arXiv preprint arXiv:1803.07416, March 2018.

[65] D.P. Kingma and J. Ba, “Adam: A Method For Stochastic Optimization,” arXiv preprint arXiv:1412.6980, Dec 2014.

[66] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens et al., “Moses: Open Source Toolkit for Statistical Machine Translation,” in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, Prague, Czech Republic, June 2007. pp. 177–180.

[67] M. Buhrmester, T. Kwang and S.D. Gosling, “Amazon’s Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?” *Perspectives on Psychological Science* vol. 6, no. 1, pp. 3–5, 2011.

[68] Z. Yang, Y. Cheng, Y. Liu and M. Sun, “Reducing Word Omission Errors in Neural Machine Translation: A contrastive learning approach,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. July 2019. pp. 6191–6196.

[69] S. Läubli, R. Sennrich and M. Volk, “Has Machine Translation Achieved Human Parity? A Case For Document-level Evaluation,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. November 2018. pp. 4791–4796.

[70] L. Guillou, C. Hardmeier, E. Lapshinova-Koltunski and S. Loáiciga. “A pronoun test suite evaluation of the English–German MT systems at WMT 2018,” in *Proceedings of the Third Conference on Machine Translation:Shared Task Papers*, Belgium, Brussels. November 2018. pp. 570–577.

[71] A. Lopes, M.A. Farajian, R. Bawden, M. Zhang, and A.F.T. Martins, “Document-level neural MT: A systematic comparison,” in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, Lisboa, Portugal, November 2020. pp. 225–234.

[72] J. Huo, C. Herold, Y. Gao, L. Dahlmann, S. Khadivi and H. Ney, “Diving deep into context-aware neural machine translation,” in *Proceedings of the Fifth Conference on Machine Translation*, Online, November 2020. pp. 604–616.

[73] J. Libovický and J. Helcl, “Attention strategies for multi-source sequence-to-sequence learning,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada, July 2017. pp. 196–202.

[74] H. Zinsmeister, S. Dipper and M. Seiss.“Abstract pronominal anaphors and label nouns in German and English: Selected case studiesand quantitative investigations,” *Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation: TC3 II*, Language Science Press, Berlin, pp. 153–195.

[75] S. Jean and K. Cho. “Context-Aware Learning for Neural Machine Translation,” arXiv preprint arXiv:1903.04715, March 2019.

[76] D. Saunders, F. Stahlberg and B. Byrne. “Using context in neural machine translation training objectives,” in *Proceedings of the 58th Meeting of the Association for Computational Linguistics*, Online, July 2020. pp. 7764–7770.

[77] D. Stojanovski and A. Fraser, “Improving anaphora resolution in neural machine translation using curriculum learning,” in *Proceedings of Machine Translation Summit XVII: Research Track*, Dublin, Ireland, August 2019. pp. 140–150.

[78] E. Lapshinova-Koltunski, M-P. Krielke and C. Hardmeier. “Coreference strategies in English-German translation,” in *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, Barcelona, Spain (online), December 2020. pp. 139–153.

[79] T. Ohtani, H. Kamigaito, M. Nagata and M. Okumura. “Context-aware neural machine translation with coreference information,” in *Proceedings of the Fourth Workshop on Discourse in Machine Translation*, Hong Kong, China, November 2019. pp. 45–50.

[80] D. Stojanovski and A. Fraser., “Coreference and coherence in neural machine translation: A study using oracle experiments,” in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium, November 2018. pp. 49–60.

[81] A. Sugiyama and N. Yoshinaga, “Data augmentation using back-translation for context-aware neural machine translation,” in *Proceedings of the Fourth Workshop on Discourse in Machine Translation*, Hong Kong, China, November 2019. pp. 35–44.

[82] G. Lample, M. Ott, A. Conneau, L. Denoyer and M. Ranzato, “Phrase-based neural unsupervised machine translation,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018. pp. 5039–5049.

[83] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning*, Online, July 2020. pp. 1597–1607.

[84] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado and J. Dean. “Distributed representations of words and phrases and their compositionality,” in *Advances in Neural Information Processing Systems 26*, Stateline, USA, December 2013. pp. 3111-3119.

[85] Z. Wu, S. Wang, J. Gu, M. Khabsa, F. Sun and H. Ma “CLEAR: Contrastive Lesarning for Sentence Representation,” arXiv preprint, arXiv:2012.15466, Dec 2020.

[86] S. Lee, D.B. Lee, and S.J. Hwang. “Contrastive Learning with Adversarial Perturbations for Conditional Text Generation,” in *Proceedings of the 8th International Conference on Learning Representations*, Online, May 2021.

[87] X. Pan, M. Wang, L. Wu and L. Li, “Contrastive learning for many-to-many multilingual neural machine translation,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, August 2021. pp. 244–258.

[88] J. Huang, Y. Li, W. Ping and L. Huang, “Large margin neural language model,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018. pp. 1183–1191.

[89] L. Yu, L. Sartran, P-S. Huang, W. Stokowiec, D. Donato, S. Srinivasan, A. Andreev, W. Ling, S. Mokra, A.D. Lago, Y. Doron, S. Young, P. Blunsom and C. Dyer, “The DeepMind Chinese–English Document Translation System at WMT2020,” in *Proceedings of the Fifth Conference on Machine Translation*, Online, July 2020. pages 326–337.

[90] L. Barrault, O. Bojar, M.R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müller, S. Pal, M. Post and M. Zampieri, “Findings of the 2019 conference on machine translation

(WMT19)," in *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, Florence, Italy, July 2019. pages 1–61.

[91] R. Sennrich, B. Haddow and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Berlin, Germany, August 2016. pages 1715–1725.

[92] M. Post. "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium, November 2018. pages 186–191.

[93] E. Lapshinova-Koltunski, C. España-Bonet and J. van Genabith, "Analysing coreference in transformer outputs," in *Proceedings of the Fourth Workshop on Discourse in Machine Translation*, HongKong, China, November 2019. pages 1–12.

[94] M. Esplà, M. Forcada, G. Ramírez-Sánchez and L. Hoang, "ParaCrawl: Web-scale parallel corpora for the languages of the EU," in *Proceedings of the MT Summit XVII*, Dublin, Ireland, August 2019. pp. 118–119.

[95] A. Viswanathan, V. Wang and A. Kononova, "Controlling Formality and Style of Machine Translation Output Using AutoML," in *Annual International Symposium on Information Management and Big Data* Lima, Peru, August 2019. pp. 306–313.

[96] Z. Xiao and A. McEnery, "Two approaches to genre analysis: Three genres in modern American English," *Journal of English Linguistics*, vol.33, pp. 62–82, 2005.

[97] E. Voita, R. Sennrich and T. Ivan, "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, Hong Kong, China, November 2019. pp. 877–886.

[98] 허광호, 고영중, 서정연, "저-자원 언어의 번역 향상을 위한 다중-언어 기계번역," *한국정보과학회 2019년 컴퓨터종합학술대회*, 제46권, 제1호, 2019년 6월. pp. 0649-0651.

[99] 정영준, 박천음, 이창기, 김준석, "MASS와 상대 위치 표현을 이용한 영어-한국어 신경망 기계 번역,'' *정보과학회논문지*, 제47권, 제11호, 1038-1043쪽, 2019년 10월.

[100] Q-P. Nguyen, D. Anh, J.C. Shin, P. Tran and C.Y. Ock, "Korean-Vietnamese Neural Machine Translation System With Korean Morphological Analysis and Word Sense Disambiguation," *IEEE Access* vol. 7, pp. 32602–32616, 2019.

[101] C. Park, Y. Yang, K. Park and H. Lim, "Decoding Strategies for Improving Low-Resource Machine Translation," *MDPI Electronics*, vol. 9, pp. 1562–1577, 2020.

[102] R. Sennrich, B. Haddow and A. Birch, "Controlling Politeness in Neural Machine Translation via Side Constraints," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, USA, June 2016. pp. 35–40.

[103] C. Chu, R. Dabre and S. Kurohashi, "An Empirical Comparison of Domain Adaptation Methods for Neural Machine Translation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, July 2017. pp. 385–391.

[104] L. Wang, M. Tu, M. Zhai, H. Wang, S. Liu and S.H. Kim, "Neural Machine Translation Strategies for Generating Honorific-style Korean," in *Proceedings*

of the 2019 International Conference on Asian Language Processing, Shanghai, China, November 2019. pp. 450–455.

[105] P. Michel and G. Neubig, "Extreme Adaptation for Personalized Neural Machine Translation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* , Melbourne, Australia, July 2018. pp. 312–318.

[106] X. Niu, S. Rao and M. Carpuat, "Multi-Task Neural Models for Translating Between Styles Within and Across Languages," in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, USA, August 2018. pp. 1008–1021.

[107] X. Niu and M. Carpuat, "Controlling Neural Machine Translation Formality with Synthetic Supervision," in *Proceedings of the AAAI Conference on Artificial Intelligence*, New York City, USA, February 2020.

[108] W. Feely, E. Hasler and A. Gispert, "Controlling Japanese Honorifics in English-to-Japanese Neural Machine Translation," in *Proceedings of the 6th Workshop on Asian Translation (WAT)*, Hong Kong, China, November 2019. pp. 45–53.

[109] A.S. Byon, "Teaching Korean Honorifics," *The Korean Language in America* vol. 5, pp. 275-289, 2000.

[110] T. Vu and G. Haffari, "Automatic Post-Editing of Machine Translation: A Neural Programmer-Interpreter Approach," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018. pp. 3048-3053.

[111] 이동주, 연종흠, 황인범, 이상구, "꼬꼬마 : 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구," *한국정보과학회논문지:컴퓨팅의 실제 및 레터*, 제16권, 제11호, 1229-7712쪽, 2020년 11월.

[112] X. Kang, Y. Zhao, J. Zhang and C. Zong, “Dynamic Context Selection for Document-level Neural Machine Translation via Reinforcement Learning,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online, November 2020. pp. 2242–2254.

[113] S. Liu and X. Zhang, “Corpora for Document-Level Neural Machine Translation,” in *Proceedings of the 12th Language Resources and Evaluation Conference,* Marseille, France, May 2020. pp. 3775–3781.

[114] L.M. Werlen and A. Popescu-Belis, “Validation of an automatic metric for the accuracy of pronoun translation (APT),” in *Proceedings of the 3rd Workshop on Discourse in Machine Translation,* Copenhagen, Denmark, September 2017. pp. 17–25.

[115] Z. Gong, M. Zhang and G. Zhou, “Document-level machine translation evaluation with gist consistency and text cohesion,” in *Proceedings of the 2nd Workshop on Discourse in Machine Translation,* Lisbon, Portugal, September 2015. pp. 33–40.

[116] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,’’ in *Proceedings of the 8th International Conference on Learning Representations*, Online, April 2020.

[117] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys,* vol. 54, no. 6, pp 1-35, 2021.

[118] E. Vanmassenhove, C. Hardmeier and A. Way, “Getting Gender Right in Neural Machine Translation,’’ in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018. pp. 3003–3008.

[119] L. Schiebinger, “Scientific research must take gender into account’’ *Nature*, vol. 507, no. 9, March 2014.

[120] B. Savoldi, M. Gaido, L. Bentivogli, M. Negri and M. Turchi, “Gender Bias in Machine Translation,’’ *Transactions of the Association for Computational Linguistics*, vol. 9 pp. 845–874, August 2021.

[121] M.R. Costa-jussà, C. Basta and G.I. Gállego, “Evaluating Gender Bias in Speech Translation,’’ arXiv preprint arXiv:2010.14465, October 2020.

[122] C. Basta, M.R. Costa-juss‘a and J.A.R. Fonollosa, “Towards mitigating gender bias in a decoder-based neural machine translation model by adding contextual information,” in *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, Seattle, USA, July 2020. pp. 99–102.

[123] G. Stanovsky, N.A. Smith and L. Zettlemoyer, “Evaluating Gender Bias in Machine Translation,’’ in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* Florence, Italy, July 2019. pp. 1679-1684.

[124] J. Troles and U. Schmid, “Extending Challenge Sets to Uncover Gender Bias in Machine Translation: Impact of Stereotypical Verbs and Adjectives,’’ in *Proceedings of the Sixth Conference on Machine Translation*, Punta Cana, Dominica republic and Online, November 2021.

[125] D. Saunders, “Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey,’’ arXiv preprint arXiv:2104.06951, April 2021.

[126] S.U. Haq, S.A. Rauf, A. Shoukat and N. Hira, “Improving Document-Level Neural Machine Translation with Domain Adaptation,’’ in *Proceedings of the Fourth Workshop on Neural Generation and Translation*, Online, July 2020. pp. 225-231.

[127] R. Chatterjee, M. Freitag, M. Negri and M. Turchi, “Findings of the WMT 2020 Shared Task on Automatic Post-Editing,’’ in *Proceedings of the Fifth Conference on Machine Translation,* Online, November 2020. pp.646-659.

[128] M. Alam, I. Kvapilíková, A. Anastasopoulos, L. Besacier, G. Dinu, M. Federico, M. Gallé, K. Jung, P. Koehn and V. Nikoulina, “Findings of the WMT Shared Task on Machine Translation Using Terminologies,’’, in *Proceedings of the Sixth Conference on Machine Translation*, Punta Cana, Dominica republic and Online, November 2021.

초 록

신경망 기계번역 기법은 최근 번역 품질에 있어서 큰 성능 향상을 이룩하여 많은 주목을 받고 있다. 그럼에도 불구하고 현재 대부분의 신경망 번역 시스템은 텍스트를 독립된 문장 단위로 번역을 수행하기 때문에 텍스트에 존재하는 문맥을 무시하고 결국 문서 단위로 파악했을 때 적절하지 않은 번역문을 생성할 수 있는 단점이 있다. 이를 극복하기 위해 주변 문장을 동시에 고려하는 문맥 인식 기반 신경망 번역 기법이 제안되고 있다. 본 학위 논문은 문맥 인식 기반 신경망 번역 시스템의 성능을 개선시킬 수 있는 기법들과 문맥 인식 기반 신경망 번역 기법의 활용 방안을 제시한다. 먼저 여러 개의 문맥 문장들을 효과적으로 모델링하기 위해 문맥 문장들을 토큰 레벨 및 문장 레벨로 단계적으로 표현하는 계층적 문맥 인코더를 제시하였다. 제시된 모델은 기존 모델들과 비교하여 가장 좋은 번역 품질을 얻으면서 동시에 학습 및 번역에 걸리는 연산 시간을 단축하였다. 두 번째로는 문맥 인식 기반 신경망 번역 모델의 학습 방법을 개선하고자 하였는데 이는 기존 연구에서는 문맥에 대한 의존 관계를 전부 활용하지 못하는 전통적인 음의 로그우도 손실함수에 의존하고 있기 때문이다. 이를 보완하기 위해 문맥 인식 기반 신경망 번역모델을 위한 상호참조에 기반한 대조학습 기법을 제시한다. 제시된 기법은 원문과 주변 문맥 문장들 사이에 존재하는 상호참조 사슬을 활용하여 대조 사례를 생성하며, 문맥 인식 기반 신경망 번역 모델들의 전반적인 번역 품질 뿐만 아니라 대명사 해결 성능도 크게 향상시켰다. 마지막으로는 맥락 정보가 필요한 한국어 경어체 번역에 있어서 문맥 인식 기반 신경망 번역 기법의 활용 방안에 대해서도 연구하였다. 이에 영어-한국어 번역 문제에 문맥 인식 기반 신경망 번역 기법을 적용하여 영어 원문에서 필수적인 맥락 정보를 추출하는 한편 한국어 번역문에서도 문맥 인식 사후편집 시스템을 활용하여 보다 일관된 한국어 경어체 표현을 번역하도록 개선하는 기법을 제시하였다.

주요어: 신경망 기계번역, 문서 단위 번역, 대조 학습, 심층 신경망
학번: 2014-21693

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my immediate advisor, Prof. Kyomin Jung for his great support, guidance, suggestions and criticism during the course of this study. Sincere thanks are also to Prof. K.S. Shim, Prof. B.T. Zhang, Prof G.H. Kim and Prof. K.W. Park for their thoughtful discussions and inexhaustible patience during the correction phase of this dissertation

I would also like to thank H.G. Yun, H.H. Lee, M.W. Lee, Dr. Y.H. Kim, Dr. J.B. Shin and Dr. S.H. Yoon for helping me with important comments and suggestions, and also thanks to all other MILAB colleagues who provided countless expertise and insight that greatly improved the research.

I am also thankful to all TEAM AMADEUS members (Min, Heetae, Boram, Moonhyeob), TAPEDECK members (especially Godspeed, ParkInStyle, BlueLink, Woney, Cheol-no and Mirage), WALKMAN FACTORY members (especially Moonrise and Gizmo) and all others who helped me to have a great hobby and made my academic year a very pleasant one.

Finally, I would express a deep sense of gratitude to my parents who have always stood by me with their constant love and encouragement. Special thanks are due to my one and only loving younger brother Yongseok who always strengthened my morale by standing by me in all situations.