applied sciences

MDPI

Article

Factors Behind the Effectiveness of an Unsupervised Neural Machine Translation System between Korean and Japanese

Yong-Seok Choi [1], Yo-Han Park [1], Seung Yun [2], Sang-Hun Kim [2] and Kong-Joo Lee [1,*]

[1] Department of Radio and Information Communications Engineering, ChungNam National University, 99 Daejak-ro, Yuseong-gu, Daejeon 34134, Korea; yongseok.choi.92@gmail.com (Y.-S.C.); happy005012@naver.com (Y.-H.P.)

[2] Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon 34129, Korea; syun@etri.re.kr (S.Y.); ksh@etri.re.kr (S.-H.K.)

* Correspondence: kjoolee@cnu.ac.kr; Tel.: +82-42-821-5662

Abstract: Korean and Japanese have different writing scripts but share the same Subject-Object-Verb (SOV) word order. In this study, we pre-train a language-generation model using a Masked Sequence-to-Sequence pre-training (MASS) method on Korean and Japanese monolingual corpora. When building the pre-trained generation model, we allow the smallest number of shared vocabularies between the two languages. Then, we build an unsupervised Neural Machine Translation (NMT) system between Korean and Japanese based on the pre-trained generation model. Despite the different writing scripts and few shared vocabularies, the unsupervised NMT system performs well compared to other pairs of languages. Our interest is in the common characteristics of both languages that make the unsupervised NMT perform so well. In this study, we propose a new method to analyze cross-attentions between a source and target language to estimate the language differences from the perspective of machine translation. We calculate cross-attention measurements between Korean–Japanese and Korean–English pairs and compare their performances and characteristics. The Korean–Japanese pair has little difference in word order and a morphological system, and thus the unsupervised NMT between Korean and Japanese can be trained well even without parallel sentences and shared vocabularies.

Keywords: MASS; pre-trained generation model; unsupervised neural machine translation; language typology; writing script; SOV word order

Citation: Choi, Y.-S.; Park, Y.-H.; Yun, S.; Kim, S.-H.; Lee, K.-J. Factors Behind the Effectiveness of an Unsupervised Neural Machine Translation System between Korean and Japanese. *Appl. Sci.* **2021**, *11*, 7662. https://doi.org/10.3390/app11167662

Academic Editor: Mauro Castelli

Received: 26 July 2021
Accepted: 19 August 2021
Published: 21 August 2021

1. Introduction

Masked Sequence-to-Sequence (MASS) [1] is a pre-training method for language generation. It adopts Transformer [2] as a basic architecture that consists of an encoder and decoder. The encoder takes a sentence with a randomly masked fragment as its input, and the decoder tries to predict this masked fragment. In this way, MASS can pre-train both the encoder and the decoder jointly using only unlabeled data, and the pre-trained encoder and decoder can be applied to most language-generation tasks, including Neural Machine Translation (NMT).

In this study, we pre-train a language-generation model using MASS with two monolingual corpora and then fine-tune the pre-trained model with the same corpora and a back-translation loss for the unsupervised NMT task. The interesting language pair of the NMT system is Korean–Japanese. Korean and Japanese use different writing scripts, but they have very similar typological properties, such as the word order of the Subject-Object-Verb (SOV) [3].

Since Korean and Japanese have different writing scripts, there are only a few shared vocabularies between the two languages. However, alphabets, digits, and some Chinese characters they share can occur in both corpora. In our preliminary experiments, therefore, only Japanese text without alphabets and digits and Korean text without Chinese characters

are gathered into the Japanese and Korean monolingual corpora, respectively. By doing this, we allow the smallest number of common vocabularies in the NMT system between the two languages. Only a few punctuations are shared by the two corpora.

We also build an unsupervised Korean–English NMT system using the same method. To compare with the Korean–Japanese NMT, the Korean–English NMT system is built to have as many common vocabularies between Korean and English as possible.

Korean, Japanese, and English do not share writing scripts; however, Korean–Japanese and Korean–English have very different characteristics. Due to the geographic proximity, Japanese and Korean share considerable similarities in typological features of their syntax and morphology, while having different writing scripts [4,5]. It is widely accepted that unsupervised NMT systems with high accuracies are possible only when the source and target language have a high lexical resemblance with the same scripts.

Although the vocabularies shared by Korean and Japanese are only a few punctuations, the unsupervised NMT system can learn to translate between the two languages without a parallel corpus. Unlike a Korean–Japanese pair, the unsupervised Korean–English NMT system shows disappointing results because the language pair has a different morphology system and word order despite having as many common vocabularies as possible. These preliminary experiments show that the similarities of morphology and syntax between the two languages can be an essential prerequisite for successfully training an unsupervised NMT without common vocabularies.

In this study, we want to discover what knowledge a model acquires through the pre-training process and what knowledge from the pre-trained model enables an unsupervised NMT with high accuracy. Therefore, we analyze which layers of the unsupervised NMT system store what kind of information, and whether cross-attentions demonstrate differences in the languages, such as word order. These findings can give us insight into how to train an unsupervised NMT system between various language pairs.

The contributions of this paper are as follows:

- We propose a new method to analyze cross-attentions of an encoder–decoder architecture considering the difference in properties between a source and target sequence.
- We apply the analysis method to an unsupervised NMT task between Korean and Japanese, which are known to be very similar each other.
- We demonstrate how similar Korean and Japanese are in terms of machine translation by examining cross-attentions of an encoder–decoder on which the unsupervised NMT between both languages is implemented.

The rest of the paper is organized as follows. We first explore related studies in Section 2, and we describe the training datasets and unsupervised NMT systems and report preliminary experimental results of the NMT systems in Section 3. We then propose a new analysis method of cross-attentions of NMT systems and present the analysis results and discussion in Section 4, followed by the conclusion in Section 5.

2. Related Studies

2.1. MASS

MASS is a method that pre-trains an encoder and decoder for language generation [1]. Unlike other pre-trained models such as BERT [6] and GPT [7], MASS adopts an encoder–decoder architecture because it pre-trains the encoder and decoder jointly. Hence, MASS pre-trains a model using sequence-to-sequence data. A source sequence for an encoder contains a special token, 'M', for a masked fragment, and the target sequence is the string for the masked fragment. MASS pre-trains the decoder to recover the masked segment given the context represented by the encoder. In this way, the model can learn not only how to represent a source sequence in an encoder, but also how to generate a target sequence in the decoder given the encoder's context.

A pre-trained model can be adopted for language-generation tasks such as NMT. It is also possible to develop unsupervised NMT systems by fine-tuning pre-trained models with a source and a target language together. Unsupervised NMT systems built in this way

are capable of bi-directional translation because MASS teaches the pre-trained models both languages at the same time.

MASS achieves significant improvements in an unsupervised machine translation over baselines without a pre-trained model.

2.2. Unsupervised NMT

Since NMT requires a large parallel corpus, an unsupervised NMT that relies solely on a monolingual corpus can be an initial solution for language pairs, including low-resource languages [8,9]. There are two approaches to training an NMT system in an unsupervised manner [10].

Denoising is the same technique as is used in denoising autoencoders [11], in which a system is trained to reconstruct the original version of a corrupted input sentence. To build an unsupervised NMT system using the denoising technique, Artetxe et al. [10] proposed an NMT system that consisted of one encoder shared by two languages and two decoders for each language. The training process was as follows: The NMT system takes an input sentence in a given language, encodes the sentence using the shared encoder, and then reconstructs the original sentence using the decoder of that language. This training process is essentially a trivial copying task. In order to make the system truly learn the compositionality of its input sentence, a random noise is inserted into the input sentence. As the system tries to recover the original input by denoising the random noise from an input, it can eventually learn about the internal structure of the languages. After completing the learning process, the system can translate an input sentence to a target sentence by replacing the decoder with that of the target language.

Back-translation is another approach to building an unsupervised NMT system [10,12]. First, an NMT system translates an input sentence into a target sentence that is decoded with a greedy approach. While the input sentence is genuine, the target sentence might be synthetic and artificial. In a training process with back-translation, the unsupervised NMT system learns to translate a synthetic translation into the original input sentence. The system can be updated by this training process and can generate more natural pseudo-parallel sentence pairs. This process is repeated until the NMT system reaches acceptable accuracy.

In this study, we adopt a back-translation approach to build an unsupervised NMT system because MASS utilizing a shared encoder and decoder for both languages can easily apply the back-translation approach. Both approaches assume that cross-lingual word embeddings can be built between two languages before training an unsupervised NMT system. Most embedding mapping methods independently learn the embedding for each language using monolingual corpora separately, and then map them onto a shared space by linear transformation and a bilingual dictionary.

In this study, we train cross-lingual word embeddings using a shared Byte Pair Encoding (BPE) [13] between the two languages. Generally speaking, monolingual corpora containing as many shared lexical entries as possible between the two languages are collected and utilized to maximize the benefit from the shared BPE of an unsupervised NMT.

3. Pre-Trained Models and Unsupervised NMTs

3.1. Datasets and Pre-Training Setup

In this work, the language pairs in question for machine translation are Korean–Japanese and Korean–English. We use the following four monolingual corpora that include 5 million sentences each, collected from news articles.

3.1.1. Monolingual Corpora

(*K*) The Korean corpus is collected from the Korean Contemporary Corpus of Written Sentences (http://nlp.kookmin.ac.kr/kcc/, accessed on 19 August 2021). To not only maximize lexical overlap with the English corpus but also minimize lexical overlap with Japanese corpus, we collect sentences including as many alphabets and digits

as possible, but not Chinese characters. They are tokenized using the ETRI parser (https://aiopen.etri.re.kr/service_api.php, accessed on 19 August 2021).

(*J*) The Japanese corpus is collected from WMT2019 (https://www.statmt.org/wmt19/translation-task.html, accessed on 19 August 2021) and JParaCrawl [14]. To minimize lexical overlap with the Korean corpus, we only collect sentences with neither alphabets nor digits. They are tokenized using the Mecab tokenizer (https://taku910.github.io/mecab/, accessed on 19 August 2021).

(*E*) The English corpus is collected from WMT 2017-2019 datasets (https://www.statmt.org/wmt17/translation-task.html;https://www.statmt.org/wmt18/translation-task.html;https://www.statmt.org/wmt19/translation-task.html, accessed on 19 August 2021), and those sentences are tokenized by the Moses tokenizer (https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl, accessed on 19 August 2021) for pre-processing.

These corpora consist of 5 million sentences each. They are all used in building the pre-trained generation models.

First, we build a monolingual lexicon for each corpus separately using the BPE algorithm [13] fixed at 60,000 sub-word units. The results are shown in the first column of Table 1. Then, we build a joint lexicon between the two languages with the same fixed number of sub-word units using the BPE algorithm. The numbers of (shared) vocabularies in the joint lexicons are shown in the second column of Table 1. Because the Japanese corpus does not contain alphabets and digits, the number of shared vocabularies of Korean–Japanese is only 33, most of which are punctuations. On the contrary, since most Korean sentences contain alphabets and digits, the ratio of the shared vocabulary of Korean–English rises to more than 30% despite the use of different writing scripts.

Table 1. The numbers of vocabularies in the lexicons built from each corpus.

Language Pair for Pre-Training	The Number of Voc. in Monolingual Lexicon	The Number of Shared Voc./The Number of Voc. in Joint Lexicon (the Ration)
Korean–Japanese	Korean: 37,670 Japanese: 37,510	33/75,147 (0.089%)
Korean–English	Korean: 47,805 English: 38,085	20,174/65,176 (30.9%)

3.1.2. Parallel Corpora

Parallel corpora are used in supervised training of NMTs. The 50,000 pairs of sentences are used for training NMT systems for each pair of languages. The Korean–English parallel corpus consists of data from the AI Hub site (https://aihub.or.kr/aidata/7974, accessed on 19 August 2021), and the Korean–Japanese parallel corpus is from [15].

For evaluation, we manually collect 2000 sentence triples of Korean–English–Japanese.

3.2. Pre-Trained Models and Unsupervised NMTs

We build a pre-trained model for each language pair using the same hyper-parameters and model configuration, as suggested in [1]. For the sake of clarity, the hyper-parameters used in this study are presented in Table 2. We shared the same hyper-parameters to build a pre-trained model, an unsupervised NMT, and a supervised NMT system except for the mini-batch size.

An unsupervised NMT system is built based on a pre-trained model using the same monolingual corpora and a back-translation loss while a supervised NMT system is built based on a pre-trained model using parallel corpora and a cross-entropy loss.

Table 2. The hyperparameters to build the pre-trained model, the unsupervised NMT, and the supervised NMT models.

Hyperparameters	Values
architecture	Transformer (encoder 6 layers, decoder 6 layers)
embedding dim.	1024
hidden layer dim.	1024
feed-forward dim.	4096
attention head	8
activation function	Gelu
optimizer	Adam
scheduler	inverse sqrt
initial learning rate	10^{-4}
beta1, beta2	0.9, 0.98
dropout	0.1
mini-batch size	3000 tokens (pre-trained model) 2000 tokens (supervised/unsupervised NMT)

Table 3 lists the preliminary results of the NMT systems, measured by BLEU scores [16].

We can observe the following three facts from the experimental results. First, pre-trained models are beneficial to improving the performance of a supervised NMT system. Second, when trained with the same number of parallel sentences and evaluated on the same sentence triples, the NMT system of Korean–Japanese performs better than that of Korean–English in every case. Third, the BLEU scores of the tasks of Korean–Japanese unsupervised NMT are quite high—32.76 for Japanese to Korean, and 29.07 for Korean to Japanese—even when there is no lexical overlap in the training dataset between the two languages. However, the Korean–English unsupervised NMT shows very disappointing results despite sharing the greatest number of vocabularies between the two languages. We can anticipate that the NMT system for the pair of Korean–English cannot be trained well without a parallel corpus.

Table 3. The performances of the neural machine translation systems.

Language Pairs	Models	BLEU Scores		
		Based on Pre-Trained Models		Supervised NMT
		Unsupervised NMT	Supervised NMT	
Korean–Japanese (KO-JA)	JA → KO	32.76	51.69	48.12
	KO → JA	29.07	41.96	41.02
Korean–English (KO-EN)	EN → KO	3.62	30.14	21.83
	KO → EN	1.10	27.13	19.49

In the following section, we list several experiments we performed to determine factors behind the effectiveness of an unsupervised NMT system between Korean and Japanese.

4. Analysis and Discussion

4.1. Analysis Objectives

The model pre-trained by MASS adopts the Transformer as a backbone architecture consisting of an encoder and a decoder. The encoder and decoder have six layers each, and the pre-trained model has cross-attention between the encoder and decoder. Therefore, we split the weights of the model into three groups. The first is the weights of word embeddings, the second is those of self-attention layers, and the third is the weights of encoder–decoder cross-attention layers. We aim to find the answers to the following questions.

1. Are word embeddings of fine-tuned NMT models located appropriately in a vector space according to their meanings?
2. Can word embeddings shift in a vector space appropriately when they are recalculated through self-attention layers of an encoder?
3. We assume that cross-attention layers between an encoder and a decoder reflect syntactical differences between the two languages in a certain way. Can we measure the language differences by analyzing cross-attentions between an encoder and a decoder?

4.2. Analysis of Cross-Word Embeddings

4.2.1. Word Translation: Cross-Lingual Alignment of Word Embeddings

To examine whether word embeddings of two languages are located closely in a vector space when they have similar meanings, we perform the following word-translation experiments.

We obtain bilingual dictionaries from Project MUSE [17]. They have 20,549 Korean–English translated word pairs, 22,357 English–Korean pairs, 25,969 Japanese-English pairs, and 35,353 English-Japanese pairs. For our experiments, we align those word pairs to make English–Korean–Japanese word triples, so we collect 10,295 triples.

We translate a source word into a target word by choosing the K-nearest neighbors ($K = 1$) to the source word in the vector space of cross-lingual word embeddings. The performances of word translation shown in Table 4 are measured by precisions, which are the ratio of correct translations out of the total words. The higher score indicates the better result.

Table 4. The performances of word translation.

Models	MASS Encoder			
	JA → KO	**KO → JA**	**EN → KO**	**KO → EN**
Unsupervised NMT	13.49	7.31	4.92	1.27
Supervised NMT	8.37	4.83	3.61	0.73

Shared vocabularies between two languages play an important role in grouping word embeddings according to their meanings, not their languages, in a vector space. Therefore, word embeddings from two different languages are not easy to map onto each other without shared vocabularies.

However, the word-translation task shows an unexpected result. The word embeddings of the Korean–Japanese unsupervised NMT system are better clustered according to their meanings than those of the Korean–English unsupervised NMT system, although the former had few shared vocabularies, while the latter had many.

As long as two languages have many common characteristics in morphology and syntax, it is possible to construct cross-lingual word embeddings, even if the two languages share little vocabulary with each other.

One interesting result is that the precisions of supervised NMTs are lower than those of unsupervised NMTs in both Korean–Japanese and Korean–English. The reason is thought to be that supervised learning shifts the embeddings of a few words contained in a parallel corpus in a vector space, taking into account the contexts in which the words are used. However, a word-translation task does not consider the contexts and only considers a neutral meaning of the word.

4.2.2. Sentence-Translation Retrieval

For experiments on sentence-translation retrieval, we use the same evaluation set described in Section 3.1, consisting of 2000 triples of aligned English–Korean–Japanese sentences.

The final embeddings of words in a query sentence can be acquired by calculating the weights of the encoder's self-attention layers, and the query sentence embedding can be

acquired by averaging those final embeddings. A sentence translation can be carried out by retrieving a target sentence whose embedding is nearest to that of a query sentence.

We use 2000 queries of Japanese and English and 200,000 Korean candidates, including 2000 target sentences. Table 5 presents the results and the performances are also estimated by precisions, which are the ratio of correctly translated sentences among the total 2000 sentences. In Table 5, 'Initial embeddings' refers to word embeddings prior to the first layer of an encoder, while 'Final embeddings' refers to word embeddings recalculated through the encoder's layers.

Table 5. The performances of sentence-translation retrievals.

Models	Embeddings	MASS Encoder	
		JA → KO	EN → KO
Unsupervised NMT	Initial embeddings	6.10	1.05
	Final embeddings	20.90	2.95
Supervised NMT	Initial embeddings	1.15	0.90
	Final embeddings	76.85	40.20

Like the word-translation task, the sentence-translation task from Japanese to Korean outperforms that from English to Korean.

In addition, as easily guessed, the precisions of the supervised NMTs are higher than those of the unsupervised NMTs. The precision of Japanese to Korean rose to 76.85% when supervised learning is adopted in the pre-trained models. The precision of sentence translation from Japanese to Korean in unsupervised learning goes up to 20.90% accuracy, although the NMT model is never trained with any parallel sentences.

The precisions of the final embeddings were much better than those of the initial embeddings in the case of Japanese to Korean, regardless of how the NMT systems are trained. This indicates that the encoder's internal layers are well trained enough to appropriately move word embeddings in a vector space, taking into account the context in which they are used.

Only final embeddings of supervised NMT seemed to work for sentence translation from English to Korean. This is analogous to the results in Table 1, in which a supervised learning process is indispensable for an English–Korean NMT system.

4.3. Analysis of Cross-Attentions

Cross-attentions can be seen as a reordering model as well as an aligning model between two languages from the perspective of machine translation [18]. Although cross-attentions are known to be a core component of deep learning networks for translation, few studies have analyzed the relationships between cross-attentions and a language pair. In this study, we analyze the properties of cross-attentions between a source language and a target language. Prior to analysis, we first define a few terminologies about attentions used in this work.

Most NMT models are implemented based on an encoder–decoder architecture. In the encoder–decoder architecture, a cross-attention mechanism is devised so that the decoder can decide which part of a source sentence to pay attention to when generating the next token. The cross-attention mechanism can relieve the encoder of the burden to represent all the information of a source sentence in a fixed-size vector.

The cross-attention of each decoding step indicates which source part should be used to translate into a target language. Therefore, in this work, we define the source token with the highest attention score at each decoding step as an aligning token for the current target token.

Assuming that a source and a target are the same, the sequence of aligning tokens for the target sequence monotonically increases, as shown in Figure 1a. In this figure, the arrow

points to an aligning token in a source sequence. We name it a *fixed cross-attention* and denote it by $A^{fixed}_{L \to L}$ from a language L to the same language L.

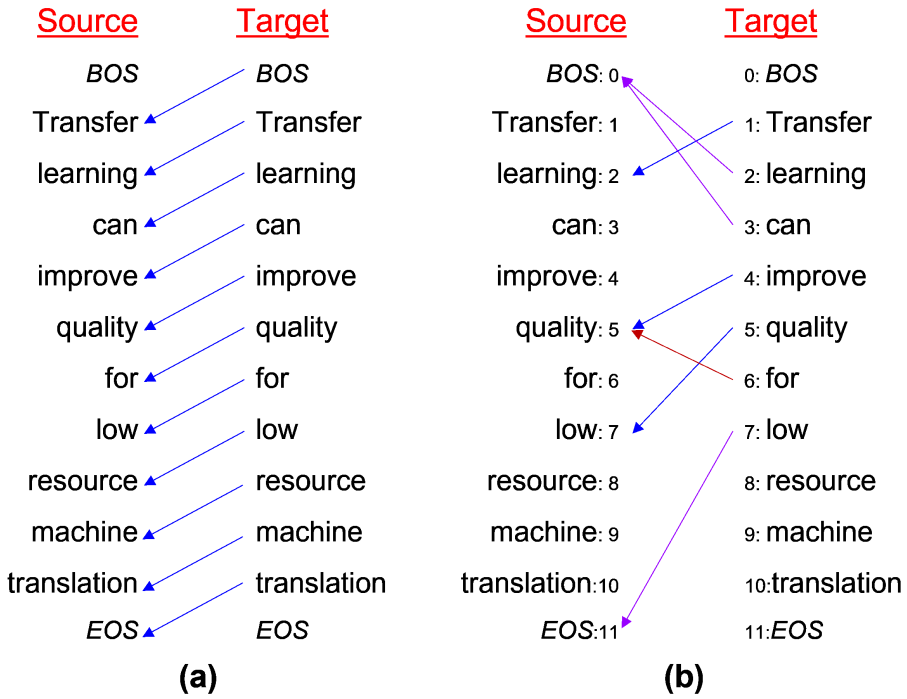

Figure 1. (**a**) Fixed cross-attention $A^{fixed}_{E \to E}$ between the same sentence with monotonic increasing. (**b**) cross-attentions that are non-aligned and aligned in the reverse order.

First, we analyze cross-attentions between the same sentences in NMT models. Since MASS can train an NMT model bidirectionally, the NMT model implemented using MASS in this study can translate sentences in both directions.

Therefore, it is possible to examine cross-attentions between the same sentences. Ideally, the cross-attentions would be nearly the same as the fixed cross-attentions because the source and the target sentences are the same in this case. However, the cross-attentions between the same sentences may be different from the fixed cross-attentions because the NMT model learns two languages together. Therefore, the other language's properties, such as word order and morphology, can affect the cross-attentions within the same language.

Let a source sequence be {BOS, $s_1, s_2, \ldots, s_n$, EOS} and a target sequence be {BOS, $t_1, t_2, \ldots, t_m$, EOS}. We align each token of a target sequence with a token of a source sequence. For each target token t_i, we can find a source token s_j that has the highest attention score and then denote the alignment by (t_i, s_j) for the target token t_i and its aligning token s_j.

In this work, a target token (except the final one) of a target sequence is said to be *non-aligned* when its aligning token in a source sequence is either 'BOS' or 'EOS'. For a given a alignment (t_i, s_j), we can find the largest index k of the alignment (t_k, s_i), such that $k < i$ and $s_l \notin$ {BOS, EOS}. When the indexes of two alignments are $l > j$, the target token t_i is said to be *aligned in the reverse order*.

In the example of Figure 1b, the target tokens t_2, t_3, and t_7 are non-aligned because they attend to 'BOS' or 'EOS'. The t_6 is the token aligned in the reverse order because there are two alignments (t_6, s_5) and (t_5, s_7) in the cross-attention.

The NMT model we built in this work has six layers in the decoder, and each layer has eight heads of cross-attention; therefore, the decoder of the NMT model has 48 cross-attention heads in total. We choose the best cross-attention head among the 48 that have the smallest number of tokens non-aligned and aligned in the reverse order.

Figure 2 shows the examples of alignments of NMT models of English–Korean and Japanese–Korean, respectively. In these examples, the alignments (from sixth to ninth tokens in Korean) between Japanese and Korean do not include non-aligned target tokens and tokens aligned in the reverse order. However, the alignments between English and Korean include the non-aligned token t_9 and the token t_8 aligned in the reverse order.

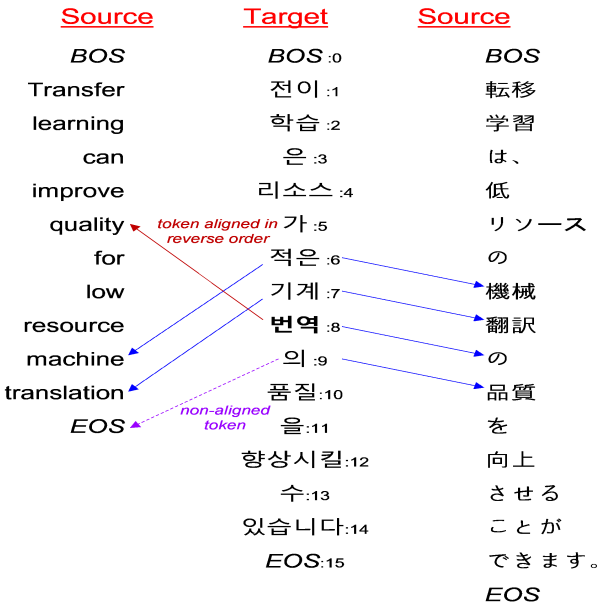

Figure 2. The examples of cross-attentions between English–Korean and Japanese–Korean.

4.3.1. Cross-Attention between the Same Sentences

The first analysis of cross-attentions, which is illustrated in Figure 3, is conducted on translation of the same sentence.

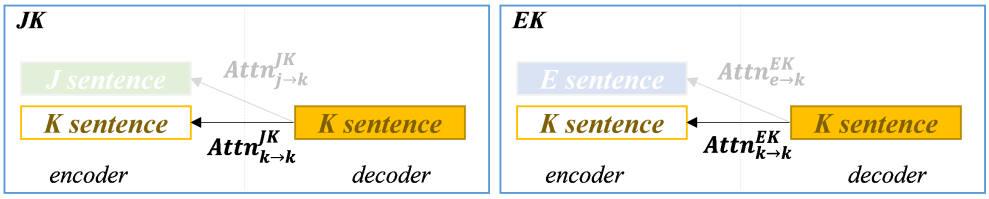

Figure 3. Comparison of cross-attentions between the same sentences.

We can easily deduce that the number of tokens aligned in the reverse order increases when the two languages differ in word order. The number of non-aligned tokens likewise increases when there are many words with different roles between the two languages. The fixed cross-attention $Attn^{fixed}_{k\to k}$ from Korean to Korean always attends to the next word; therefore, the number of tokens aligned in the reverse order and non-aligned tokens will be zero. The notation $Attn^{JK}_{k\to k}$ indicates a cross-attention provided by an NMT model when the NMT model, which can translate Japanese and Korean bi-directionally, translates a Korean sentence into the same Korean sentence.

If Korean and Japanese are similar, $Attn^{JK}_{k\to k}$ is nearly the same as the fixed cross-attention $Attn^{fixed}_{k\to k}$. However, if the two languages are quite different, the numbers of tokens aligned in the reverse order and non-aligned tokens will increase.

For 2000 Korean sentences, we analyze cross-attentions of the NMT models and count the number of tokens aligned in the reverse order and non-aligned tokens for Japanese to Korean and English to Korean, respectively. The results are shown in Table 6 and Figure 4.

Table 6. The cross-attentions between the same sentences in Japanese–Korean and English–Korean neural machine translation (NMT) models.

Cross Attentions	Models	$Attn^{JK}_{K\to K}$	$Attn^{EK}_{K\to K}$
Aligned in the reverse order	Pre-trained Model	0.204%	3.715%
	Unsupervised NMT	0.109%	0.527%
	Supervised NMT	0.514%	3.825%
Non-aligned	Pre-trained Model	0.157%	4.808%
	Unsupervised NMT	8.599%	6.789%
	Supervised NMT	10.346%	18.366%
Aligned in the reverse order + Non-aligned	Pre-trained Model	0.691%	8.523%
	Unsupervised NMT	9.447%	8.353%
	Supervised NMT	11.955%	30.874%

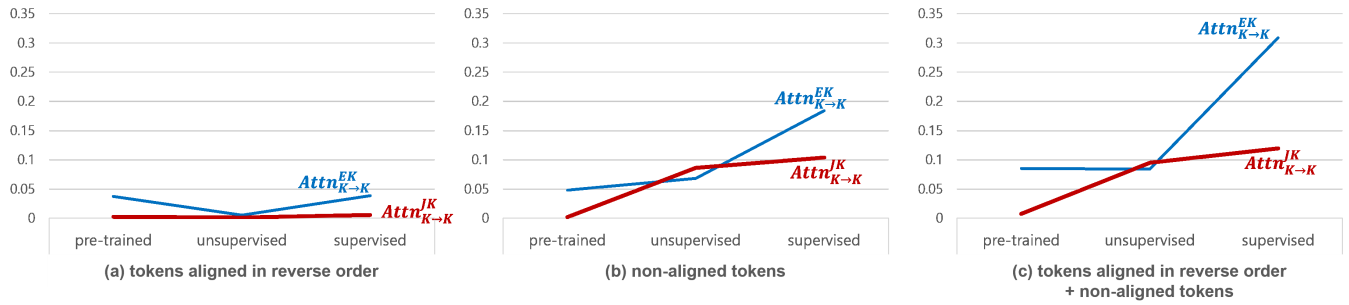

Figure 4. The cross-attentions between the same sentences in Japanese–Korean and English–Korean neural machine translation (NMT) models.

First, the ratios of tokens non-aligned and aligned in the reverse order in $Attn^{JK}_{k\to k}$ and $Attn^{EK}_{K\to K}$ are quite distinguishable. While 11.955% of the Japanese–Korean model's tokens are non-aligned or aligned in the reverse order, 30.874% of the English–Korean model's tokens are non-aligned or aligned in the reverse order, which is more than twice that of the Japanese–Korean model.

In addition, the results of Figure 4 show that the cross-attentions of the Japanese–Korean model change significantly after unsupervised learning, while those of the English–Korean model change significantly after supervised learning. In other words, Japanese–Korean models can learn the properties of the two languages with unsupervised learning only; however, English–Korean models hardly learn the properties of the two languages without supervised learning.

From the fact that the ratio of tokens aligned in the reverse order in the Japanese–Korean model is far below 1%, we can assert that Korean and Japanese are quite similar in word order.

4.3.2. Cross-Attention between the Different Languages

This analysis is designed to observe how the cross-attentions are calculated when a source and a target language are quite different. This analysis is illustrated in Figure 5. The results are presented in Table 7 and Figure 6.

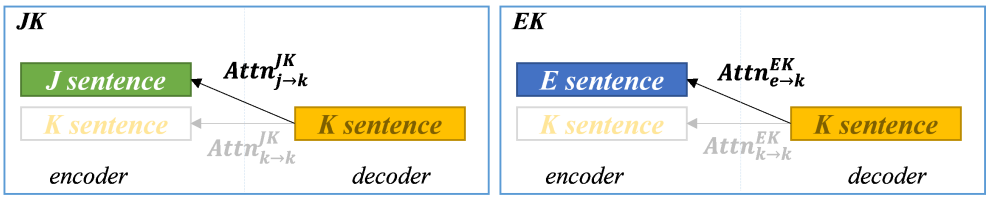

Figure 5. Comparison of cross-attentions between the different languages.

Table 7. The cross-attentions between Japanese–Korean and English–Korean.

Cross Attentions	Models	$Attn^{JK}_{J\to K}$	$Attn^{EK}_{E\to K}$
Aligned in the reverse order	Pre-trained Model	0.992%	3.284%
	Unsupervised NMT	0.703%	2.069%
	Supervised NMT	0.992%	7.256%
Non-aligned	Pre-trained Model	2.112%	8.126%
	Unsupervised NMT	10.078%	18.917%
	Supervised NMT	12.099%	16.095%
Aligned in the reverse order + Non-aligned	Pre-trained Model	4.446%	14.202%
	Unsupervised NMT	12.055%	25.027%
	Supervised NMT	14.651%	41.279%

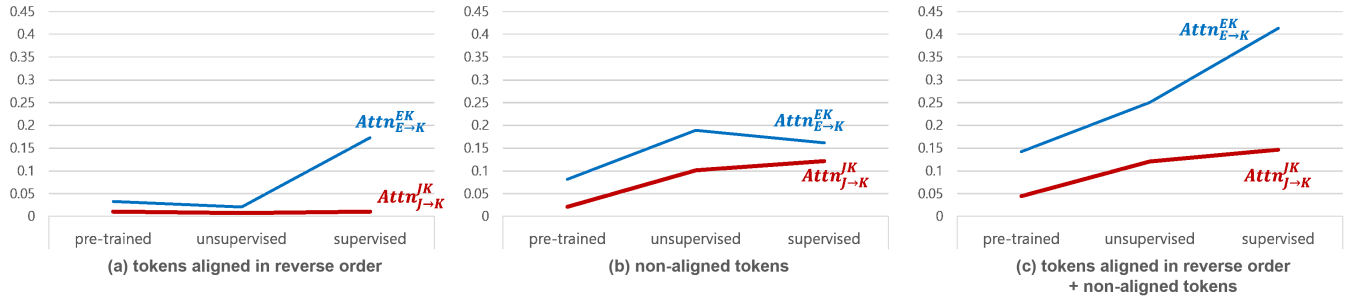

Figure 6. The cross-attentions between Japanese–Korean and English–Korean.

(1) We can assert that the respective word orders of Korean and Japanese are almost the same. In the attention $Attn^{JK}_{J\to K}$ of the Japanese–Korean model, the number of tokens aligned in the reverse order is still below 1%, while that of the English–Korean model soars to 17.256%.

(2) The number of non-aligned and aligned tokens in the reverse order in $Attn^{EK}_{E\to K}$ goes up to 41.279%, while that in $Attn^{JK}_{J\to K}$ still remains at 14.651%.

From the analyses of cross-attentions between the languages, we can conclude that Japanese and Korean are very similar in terms of machine translation. The word orders of the two languages are very similar, and thus the alignment between the two languages monotonically increases, excepting about 1%, and below 15% when considering non-aligned tokens. Conversely, about 41% of the alignments of cross-attentions between English and Korean are reversed in word order, or has meaningless attentions.

5. Conclusions

In this study, we built unsupervised NMT systems for pairs of Korean–Japanese and Korean–English and analyzed the performances of the NMT systems according to the properties of the language pairs.

We proposed a new method to analyze cross-attentions in terms of alignments. Using this analysis method, we were able to draw insights about the differences in word order between a source and a target language in an NMT model.

Korean and Japanese are known to be very similar in word order, morphology, and syntax; however, there has been no way to measure their similarities. In this study, we measured the similarities between the two languages by alignments of cross-attentions. The high similarity between the two languages allows unsupervised NMT systems of the languages to achieve high performance.

Before building an unsupervised NMT system, we can estimate the differences between a language pair using the proposed method, and thereby predict whether an unsupervised NMT system between the language pair can be trained well.

Author Contributions: Conceptualisation, S.Y., S.-H.K. and K.-J.L.; methodology, Y.-S.C. and K.-J.L.; software, Y.-H.P.; validation, Y.-H.P.; writing—original draft preparation, K.-J.L. and Y.-S.C.; writing—review and editing, K.-J.L. and Y.-S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Electronics and Telecommunications Research Institute(ETRI) grant funded by the Korean government. [21ZS1100, Core Technology Research for Self-Improving Integrated Artificial Intelligence System].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analysed during this study are included in this published article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BERT	Bidirectional Encoder Representations from Transformers
BOS	Begin of Sentence
BPE	Byte Pair Encoding
EOS	End of Sentence
GPT	Generative Pre-trained Transformer
MASS	Masked Sequence-to-Sequence
MUSE	Multilingual Unsupervised and Supervised Embeddings
NMT	Neural Machine Translation

References

1. Song, K.; Tan, X.; Qin, T.; Lu, J.; Liu, T.Y. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 5926–5936.
2. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.U.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
3. Han, J.Y.; Oh, T.H.; Jin, L.; Kim, H. Annotation Issues in Universal Dependencies for Korean and Japanese. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*; Association for Computational Linguistics: Barcelona, Spain, 2020; pp. 99–108.
4. Brown, L.; Yeon, J. *The Handbook of Korean Linguistics*; Wiley-Blackwell: Hoboken, NJ, USA, 2015.
5. Murphy, E. *The Oscillatory Nature of Language*; Cambridge University Press: Cambridge, UK, 2020. [CrossRef]
6. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4171–4186. [CrossRef]
7. Radford, A.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. *arXiv* **2018**, arXiv:1802.05365.
8. Zhang, W.; Li, X.; Yang, Y.; Dong, R. Pre-Training on Mixed Data for Low-Resource Neural Machine Translation. *Information* **2021**, *12*, 133. [CrossRef]
9. Zhang, W.; Li, X.; Yang, Y.; Dong, R.; Luo, G. Keeping Models Consistent between Pretraining and Translation for Low-Resource Neural Machine Translation. *Future Internet* **2020**, *12*, 215. [CrossRef]
10. Artetxe, M.; Labaka, G.; Agirre, E.; Cho, K. Unsupervised Neural Machine Translation. *arXiv* **2017**, arXiv:1710.11041.
11. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
12. Sennrich, R.; Haddow, B.; Birch, A. Improving Neural Machine Translation Models with Monolingual Data. *arXiv* **2015**, arXiv:1511.06709.
13. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. *arXiv* **2015**, arXiv:1508.07909
14. Morishita, M.; Suzuki, J.; Nagata, M. JParaCrawl: A Large Scale Web-Based English-Japanese Parallel Corpus. *arXiv* **2019**, arXiv:1911.10668.
15. Lee, K.J.; Lee, S.; Kim, J.E. A Bidirectional Korean–Japanese Statistical Machine Translation System by Using MOSES. *J. Adv. Mar. Eng. Technol.* **2012**, *36*, 683–693. [CrossRef]

16. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.
17. Conneau, A.; Lample, G.; Ranzato, M.; Denoyer, L.; Jégou, H. Word Translation Without Parallel Data. *arXiv* **2017**, arXiv:1710.04087.
18. Liu, L.; Utiyama, M.; Finch, A.M.; Sumita, E. Neural Machine Translation with Supervised Attention. *arXiv* **2016**, arXiv:1609.04186.