electronics

MDPI

Article

Decoding Strategies for Improving Low-Resource Machine Translation

Chanjun Park [1], Yeongwook Yang [2], Kinam Park [3] and Heuiseok Lim [1,*]

[1] Department of Computer Science and Engineering, Korea University, Seoul 02841, Korea; bcj1210@naver.com
[2] Center for Educational Technology, Institute of Education, the University of Tartu, 50090 Tartu, Estonia; yeongwook.yang@gmail.com
[3] Creative Information and Computer Institute, Korea University, Seoul 02841, Korea; spknn@korea.ac.kr
* Correspondence: limhseok@korea.ac.kr

Received: 23 July 2020; Accepted: 22 September 2020; Published: 24 September 2020

Abstract: Pre-processing and post-processing are significant aspects of natural language processing (NLP) application software. Pre-processing in neural machine translation (NMT) includes subword tokenization to alleviate the problem of unknown words, parallel corpus filtering that only filters data suitable for training, and data augmentation to ensure that the corpus contains sufficient content. Post-processing includes automatic post editing and the application of various strategies during decoding in the translation process. Most recent NLP researches are based on the Pretrain-Finetuning Approach (PFA). However, when small and medium-sized organizations with insufficient hardware attempt to provide NLP services, throughput and memory problems often occur. These difficulties increase when utilizing PFA to process low-resource languages, as PFA requires large amounts of data, and the data for low-resource languages are often insufficient. Utilizing the current research premise that NMT model performance can be enhanced through various pre-processing and post-processing strategies without changing the model, we applied various decoding strategies to Korean–English NMT, which relies on a low-resource language pair. Through comparative experiments, we proved that translation performance could be enhanced without changes to the model. We experimentally examined how performance changed in response to beam size changes and n-gram blocking, and whether performance was enhanced when a length penalty was applied. The results showed that various decoding strategies enhance the performance and compare well with previous Korean–English NMT approaches. Therefore, the proposed methodology can improve the performance of NMT models, without the use of PFA; this presents a new perspective for improving machine translation performance.

Keywords: neural machine translation; Korean–English neural machine translation; transformer; efficiency processing; post-processing; decoding strategies

1. Introduction

Natural language processing (NLP) is a subfield of artificial intelligence in which computers analyze human languages. In general, NLP is divided into three main categories: rules-based, statistics-based, and deep learning-based. In rules-based and statistics-based NLP application software, system performance is dependent on the performance of various subcomponents such as the speech tagger, syntactic parser, and semantic analyzer. In contrast, deep learning-based NLP application software are operated in an end-to-end manner, and the performance of a model is independent of the subcomponents. The processes required for each step of an end-to-end process are handled simultaneously during training. Deep learning-based NLP application software have exhibited good innovative performance in various NLP fields such as machine translation, speech recognition, and text

summarization. Numerous organizations have used deep learning-based NLP application software to transform existing rules-based and statistics-based services into deep learning-based services.

The Pretrain-Finetuning Approach (PFA) has been widely used in recent NLP research [1–5]. PFA requires a wide range of model sizes and numerous parameters; therefore, organizations attempting to utilize PFA on their existing equipment often experience slow processing speeds and memory shortages. Accordingly, organizations with insufficient servers or GPUs for deep learning may face difficulties using the latest versions of PFA to improve machine translation results. Researchers are aiming to enhance model performance through various pre-processing and post-processing strategies, to address these limitations without changing the model. For NLP tasks that must utilize low-resource language pairs, the importance of this research is even more significant. PFA requires large quantities of data, and such data are lacking for low-resource language pairs.

Recently, the number of studies conducted with a focus on NLP has rapidly increased, particularly in the neural machine translation (NMT) field [6–14]. Examples of pre-processing in NMT include subword tokenization [8,9], parallel corpus filtering (PCF) [10], alignment [11], back translation [12], and copied translation [13], which are all data augmentation techniques. Post-processing includes automatic post editing (APE) [14] and the application of various strategies such as beam size changes, n-gram blocking, and length penalties [15] during decoding. Some of these studies aimed to enhance performance through pre- and post-processing by changing only the subword tokenization without changing the model, whereas in others, performance was enhanced by applying data augmentation techniques [6,7]. The importance of PCF was emphasized by a translation model designed for Facebook; the model won the Conference on Machine Translation (WMT) 2019 [16].

NMT research on the Korean language, which is considered a low-resource language, follows a similar path. However, most Korean NMT studies have been conducted with a focus on pre-processing and models [6–9,12]; to the best of our knowledge, studies that focus only on post-processing do not exist. Therefore, through various experiments, we aimed to prove that NMT performance could be enhanced through various decoding strategies without changing the model; this is achieved using the Transformer [17] model, which achieved the best performance up till the advent of the PFA-based model. Additionally, we focused on Korean–English NMT research in this study; as noted previously, it is difficult to apply the latest versions of PFA to such low-resource language pairs.

The various decoding strategies used in the experiments in this study can be divided into three categories. Experiments analyzed how performance changed in response to beam size changes and the use of n-gram blocking, and how performance changed when the length penalty and stepwise penalty were applied. The optimal performance level was determined by applying these strategies in a gradual pipelining form rather than independently. A total of 1.6 million Korean–English sentences obtained from the NIA (http://aihub.or.kr/) and OpenSubtitles (http://opus.nlpl.eu/OpenSubtitles-v2018.php) were used as training data in the experiment. The test sets used in existing research, IWSLT-16 [18] and IWSLT-17 [19], were used as the test sets in a performance comparison. The contributions of this study are as follows.

- It is proved that performance can be enhanced through various decoding strategies without changing the model. This finding may serve as a basis for pre-processing and post-processing research in the field of NMT. This study presents a methodology that can improve performance without using PFA, and it presents a new perspective for improving machine translation (MT) performance.
- An in-depth comparative analysis applies various decoding strategies to existing Korean–English NMT approaches. To the best of our knowledge, there are no studies conducted on Korean–English NMT that compare and analyze various decoding strategies. In this study, a performance comparison is made using existing Korean–English NMT research, and the objective test sets IWSLT-16 and IWSLT-17 are used as the criteria. The decoding strategies were applied in a pipelining form. We identified the optimal beam size through a comparison of beam sizes,

applied n-gram blocking based on the optimal beam size, and applied the length penalty based on the optimal n. This gradual decoding process represents a novel approach.
- The distribution of the model was enhanced by creating it in the form of a platform. This approach contributes to the removal of lingual barriers as more people adopt the model.

This paper is structured as follows: Section 2 reviews related work in the area of Korean–English NMT and provides the background to help an understanding of our proposed approach. Section 3 discusses the proposed approach, and Section 4 describes the experiments and results. Section 5 concludes the paper.

2. Related Works and Background

2.1. Machine Translation

MT refers to the computerized translation of a source sentence into a target sentence; the advent of deep learning has significantly enhanced the performance of MT. Yehoshua-Bar-Hillel began research on MT in 1951 at MIT [20]; since then, it has developed in the following order: rules-based, statistics-based, and deep learning-based MT.

Rules-based MT (RBMT) [21,22] performs translation on the basis of traditional NLP processes such as lexical analysis, syntax analysis, and semantic analysis, in conjunction with linguistic rules established by linguists. For example, in Korean–English NMT, this methodology accepts a sentence in Korean (the source language) as input, guides it through a process of morphological and syntactic analysis, and produces output that complies with the grammatical rules of English (the target language). This methodology can produce a perfect translation for sentences that fit established rules. However, it is difficult to extract certain grammatical rules, for which significant linguistic knowledge is needed. In addition, it is difficult to expand the translation language and system complexity is relatively high.

Statistics-based MT (SMT) [23,24] is a method that uses statistical information that is trained from a large-scale parallel corpus. Specifically, it uses statistical information to perform translations on the basis of the alignment and co-occurrences between words in a large-scale parallel corpus. SMT is composed of a translation model and a language model. It extracts the alignments of the source sentence and target sentence through the translation model and predicts the probability of the target sentence through the language model. Unlike RBMT, this methodology can be developed without linguistic knowledge, and the quality of the translation improves as the quantity of data increases. However, it is difficult to obtain a large amount of data, and it can be a challenge to understand context because the translation is conducted in units of words and phrases.

The NMT method performs translation via deep learning. It vectorizes the source sentence through the encoder using the sequence-to-sequence (S2S) model, decodes the vector through the decoder, and creates a sentence in the target language. It identifies the most suitable expression and translation outcome by using deep learning and considering the input and output sentences as a pair. In other words, this methodology comprises an encoder and a decoder. It vectorizes the source sentence through the encoder, condenses the information as a context vector, and generates the translated target sentence in the decoder based on the condensed information. Methodologies utilized for NMT include recurrent neural networks (RNNs) [25,26], convolutional neural networks (CNNs) [27,28], and the Transformer model [17]. The Transformer model has exhibited better performance than the other approaches. Further, the recent trend is to apply PFA techniques such as cross-lingual language model pre-training (XLM) [29], masked sequence-to-sequence pre-training (MASS) [30], and the multilingual bidirectional and auto-regressive transformer (mBART) [31] to NMT as well, and such strategies are currently providing the best performance. However, because PFA requires numerous parameters and large model sizes, it is still impractical for organizations to apply the strategies. Therefore, the Transformer model was selected as the optimal NMT model after considering factors

such as performance, speed, and memory requirements that have been reported in previous research; experiments were then conducted based on this model.

2.2. Pre- and Post-Processing in Neural Machine Translation

In deep learning, pre-processing involves actions such as the refining, transformation, augmentation, and filtering of a model to ensure better performance before training it. Post-processing involves transforming the results that the model predicted into a better form after training.

Many studies on the NMT model have focused on improving pre-processing. Subword tokenization research aims at the resolution of unknown issues and involves splitting the entered NMT sentence into certain units. It addresses the out-of-vocabulary (OOV) problem and is a step that must be passed during the NMT pre-processing stage. Representative examples include byte pair encoding (BPE) [8] and SentencePiece (SP) [9]; such methodologies have become necessary pre-processing operations in most NMT studies. Further, research on data augmentation has been conducted based on the fact that NMT requires a substantial amount of training data. Models with good performance, such as the Transformer model, improved their learning by utilizing numerous parallel corpora, although the construction of such corpora is expensive and time-consuming. Techniques for creating a pseudo-corpus were investigated to overcome this problem. The quantity of training data can be increased by transforming a monolingual corpus into a pseudo-parallel corpus (PPC) when these methodologies are used. Representative examples include back translation [12] and copied translation [13]. Back translation creates a PPC by using an existing trained opposite-direction translator and improving the translation of the monolingual corpus; the new PPC is then added to the existing parallel corpus for use during training. In other words, it is based on the fact that a bidirectional translation model can be created with one parallel corpus during the creation of the translator, which is the advantage of NMT. In contrast, the copied translation methodology utilizes only the monolingual corpus without employing an opposite-direction translator. It is a methodology that trains by inserting the same data into the source and the target. Research focused on parallel corpus filtering (PCF) [10] aims to increase the performance of models by only using high-quality training data during training. If data obtained from the Internet is used as training data, a substantial amount of the data will be noisy data, and it is prohibitively difficult for humans to verify such data. Therefore, PCF aims to remove data that are not suitable as training data.

First, automatic post editing (APE), a type of post-processing [14], is a subfield of MT that aims to create high-quality translation results by using the APE model to automatically edit the NMT results, which is then compared with the translation results produced by an existing model. This does not imply that the NMT model itself is changed, but it is a research effort to create another NMT model that corrects the translation results of the original NMT model. Moreover, there are various decoding strategies. In the decoding task to generate a translation in NMT, decoding strategies such as n-gram blocking and length penalties are applied instead of simply generating the translation through the beam search process. Through this approach, the quality of the results predicted by the NMT model can be enhanced without changing the model structure.

Recently, a significant amount of research has been conducted on APE through WMT Shared Task and other venues. However, minimal research is being conducted on the various decoding strategies. Accordingly, we proved that the performance of Korean–English NMT could be enhanced through a comparative experiment on the various decoding strategies.

2.3. Korean–English Neural Machine Translation Research

MT-related research and services are currently active in Korea. Along with the Papago service that is being provided by Naver, organizations such as the Electronics and Telecommunications Research Institute (ETRI), Kakao, SYSTRAN, LLsoLLu, and Genie Talk of Hancom Interfree are providing MT services. In addition, research on Korean MT is being conducted by a partnership between Flitto,

Evertran, and Saltlux; these organizations constructed a Korean–English parallel corpus that was recently published in the AI Hub.

Kangwon National University [32] was the first to conduct Korean–English NMT research by applying MASS [30]. MASS is a pre-training technique that randomly designates K tokens in the input, masks them, and trains the prediction of the masked tokens. Because the tokens that were not masked in the encoder are masked in the decoder, the decoder must predict the tokens that were masked by referring only to the hidden representation and attention information provided by the encoder. This provides an environment in which the encoder and decoder can proceed through pre-training together. K denotes the number of tokens that are masked. When K is 1, one token is masked in the encoder and the decoder predicts one token that went through masking. This creates the same effect as the masked LM of BERT [1]. When K is equal to m, which denotes the total length of the entered sentence, every token on the encoder side is masked; this creates the same effect as the standard LM of GPT [2]. Kangwon National University achieved high performance by applying this strategy to Korean–English NMT.

Sogang University proposed a methodology in which large quantities of out-domain parallel corpora and multilingual corpora are applied to the NMT model training process to compensate for the lack of a bulk parallel corpus [33]. In addition, studies on expansion of the Korean–English parallel corpus through the use of back translation [34] and on Korean–English NMT using BPE [8] have also been conducted [35]. However, although the experimental results in those studies are provided in the form of BLEU scores (from 1 through 4), it is difficult to compare the performance objectively because the final BLEU scores [36] are not revealed.

Korea University (KU) routinely conducts research on NMT pre-processing. They suggested two-stage subword tokenization (morphological segmentation + SentencePiece unigram), which is a subword tokenization specialized for the Korean language, and presented a paper that described applying PCF to Korean–English NMT [6]. They also proposed a methodology that conducts training with a relative ratio when composing batches, rather than simply applying back translation and copied translation when applying the data augmentation [7]. Through this strategy, performance was higher than that achieved through the simple application of back translation. In addition, they illustrated the importance of data by conducting a study in which the AI Hub corpus, published by NIA, was applied to the Transformer model and surpassed the performance achieved by previous Korean–English NMT research [37]. In conclusion, although research on the NMT model is important, KU conducted various experiments and studies illustrating the importance of pre-processing. Moreover, the model presented in their paper was made available through a platform, which increased its distribution. The platform was selected as a "Best practice of data utilization for NIA artificial intelligence training data" (http://aihub.or.kr/node/4525).

3. Proposed Approach

3.1. Models

As a follow-up to the paper "Machine Translation Performance Improvement Research through the Usage of Public Korean–English Parallel Corpus", published by KU [37], in this study, we produced an NMT model by using the same training data, test set, and model structure used by the Transformer [17]. A performance comparison experiment was conducted by applying various decoding strategies. The overall model architecture is shown in Figure 1.

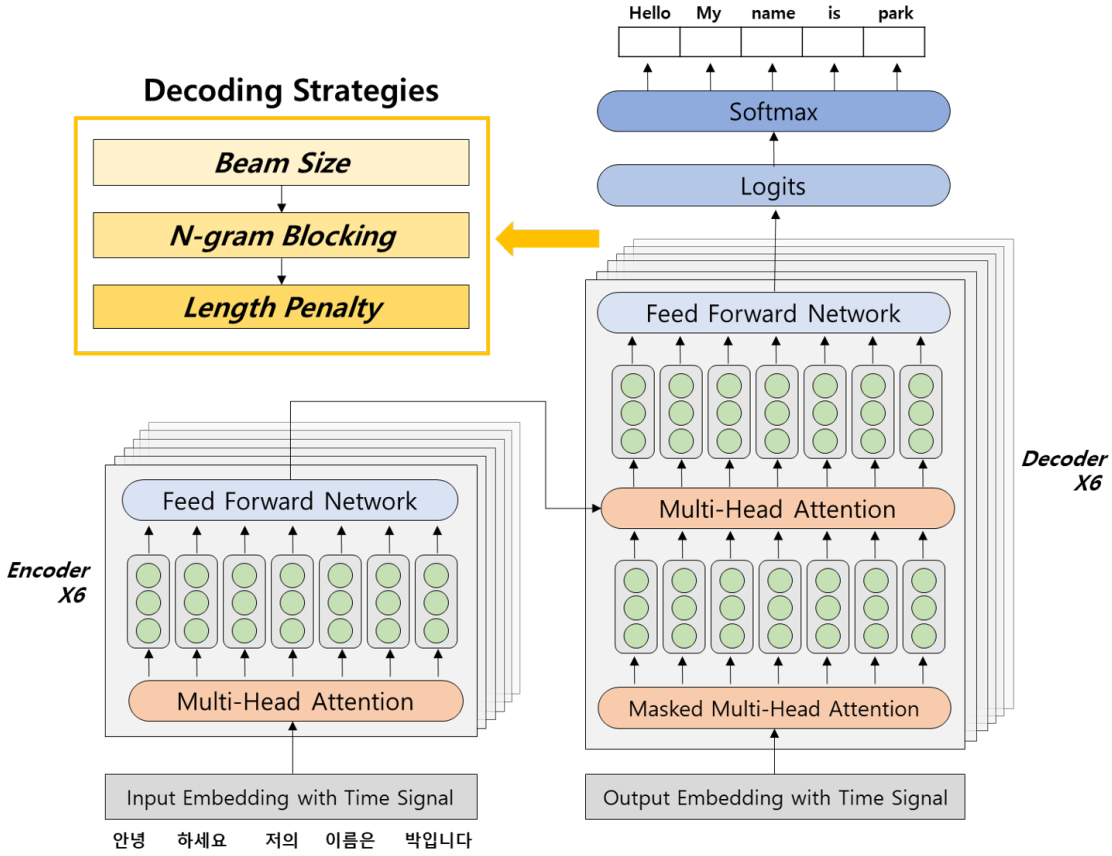

Figure 1. Overall architecture of the Korea University (KU) model using various decoding strategies.

The Transformer is a methodology that only uses attention without convolution or recurrence. Because the Transformer does not receive word entries in serial order, it adds the positional information to the embedding vector of each word to learn the words positional information and uses this as the models input. This is referred to as positional encoding. Positional encoding utilizing *sine* and *cosine* functions was added to the Transformer, and the encoding of each position is performed as follows:

$$\begin{aligned} PE_{(pos,2i)} &= \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \\ PE_{(\text{pos},\,2i+1)} &= \cos\left(\text{pos}/10000^{2i/d_{\text{model}}}\right) \end{aligned} \tag{1}$$

where pos indicates the position, and i indicates the dimension. In other words, each dimension of the positional encoding accords with a sinusoid function, and the wavelengths form a geometric progression in the range between 2π and $20{,}000\pi$. These functions were chosen because of the theoretical presumption that they would assist the model in easily learning to attend to relative positions, owing to the fact that for any fixed offset k, $PE_{pos} + k$ can be represented as a linear function of PE_{pos}.

This model is a structure that learns the attention between the input and output after learning each self-attention of the input and output based on the query, key, and value. The attention weights are then calculated using the following method:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Attention can be described as mapping a query and a set of key-value pairs to an output. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key [17]. We compute the

dot products of the query with all keys, divide each by $\sqrt{d_k}$, and apply a SoftMax function to obtain the weights on the values.

Because computational parallelism is possible, the training speed is faster than that of other models, and it is currently exhibiting good performance in the MT field. Numerous organizations are providing MT services based on this model.

In conclusion, we conducted an experiment that compared various decoding strategies on the basis of the KU model, which was in turn based on the Transformer.

3.2. Decoding Strategies

We applied various decoding strategies in the form of pipelining. First, we found the optimal beam size by comparing the performance achieved using different beam sizes and applied n-gram blocking based on this beam size. Subsequently, we found the optimal n through a performance comparison experiment and applied the length penalty.

3.2.1. Beam Search and Beam Size

Beam search is a method that increases calculation efficiency by limiting the number of candidates to be remembered in the decoding process of NMT into K; K eventually indicates the beam width or beam size. Performance is optimized when the case with the highest cumulative probability is chosen after considering every possible case; however, considering its inherent time complexity and speed, practical utilization of this method is almost impossible.

Greedy decoding is a methodology that considers every possible case; it executes a translation by simply continuing to select candidates until the candidate with the highest probability in that step appears. However, even a single wrong prediction can exert a fatal effect on overall performance; such an effect cannot be corrected even by optimal performance in subsequent steps. The beam search method was designed to overcome this problem. Therefore, beam search could be the optimal alternative for greedy decoding and the method of considering every possible case. Most NMT processes use beam search to perform decoding. Accordingly, in this study, the beam size was incremented from 1 to 10, and the resulting performance impact was examined through an experiment. Algorithm 1 shows the details of the beam search process.

Algorithm 1 Beam Search.

set Beamsize = Z;
$h_0 \Leftarrow Transformer - Encoder(S)$
$t \Leftarrow 1$
// L_S means length of source sentence;
// α is Length factor;
while $n \leq \alpha * L_S$ **do**
  $y_{1,i} \Leftarrow < EOS >$
  while $i \leq Z$ **do**
    set $h_t \Leftarrow Transformer - Decoder(h_{t-1}, y_{t,i})$;
    set $P_{t,i} = Softmax(y_{t,i})$;
    set $y_{t+1,i} \Leftarrow argTop\_Z(P_{t,i})$;
    set $i = i + 1$
  end while
  set $i = 0$
  if $h_t == < EOS >$ **then**
    break;
  end if
  set $t = t + 1$
end while
select $argmax(p(Y))$ from Z candidates Y_i
return Y_i

Beam search can be defined as the retainment of the top-k possible translations as candidates at each time step, where *k* refers to the beam width. A new possible translation is created in the next time step by combining each candidate word and a new word. The new candidate translations then compete with each other in log probability to produce the new top-*k* most reasonable results. The process continues until the translation ends.

3.2.2. N-Gram Blocking

After the advent of NMT, problems that did not appear in the existing rules-based and statistics-based methodologies started to occur. Representative examples include repeated translation, unknown (UNK) words, and omissions; these problems are generally caused by limited vocabulary size. Various subword tokenization methodologies such as BPE [8], SentencePiece [9], and BPE-Dropout [38] were suggested to solve these problems. These methodologies, which separate words into meaningful subwords, helped mitigate problems resulting from limited vocabulary size.

We determined that problems caused by repeated translation, unknown (UNK) words, and omissions can be solved through not only subword tokenization but also through n-gram blocking during decoding. Blocking improved NMT performance when a repetition occurred during decoding from unigram to 10-gram. Specifically, the output of the decoder contains the top-k beams produced through the beam search algorithm during the decoding step. At this moment, n-gram blocking is performed by referring to the output words of the previous time step, to prevent n-gram repetition and the ignoring of beams that have the same n-gram.

3.2.3. Length Penalty

The length penalty is a methodology that prevents the probability value from decreasing when a long sentence is translated. The cumulative probability is eventually calculated during the beam search, and it ranges from 0 to 1. Therefore, according to the principle of cumulative probability, the value moves closer to 0 as the translation progresses. Specifically, an unfairness occurs in which the cumulative probability eventually becomes smaller than that of a short sentence as the beam length increases. The length penalty was proposed to solve this problem. In other words, because the probability value of a long sentence necessarily decreases, the length penalty could be defined as a means of recalibrating the probability value to the sentence length. During the decoding step, we applied the stepwise length penalty (SLP) and average length penalty (ALP), which are representative length penalty methodologies. SLP implies that the penalty is applied at each decoding step, and ALP normalizes the scores based on the sequence length. Through this strategy, repeated translations can be handled more efficiently in the decoding step.

3.3. Korean–English Neural Machine Translation Platform

A distribution method was developed for the model; this was considered a crucial task. Efficient distribution of the model can assist human translators by reducing expenses and development time for Korean–English translation research. Therefore, we distributed the model as a platform to aid the development of Korean–English neural machine translation. The platform is designed to utilize both the GPU and CPU. GPUs are capable of translating at a rapid pace, thus allowing a greater number of users to experience Korean–English translation. It was distributed officially at (http://nlplab.iptime.org:32296/). The execution process of the platform is shown in Figure 2.

Figure 2. Korean–English neural machine translation platform execution process.

4. Experiment and Results

4.1. Data Sets and Data Pre-Processing

We utilized the Korean–English parallel corpus provided by the AI Hub and the Korean–English subtitles parallel corpus from OpenSubtitles as the training data for the experiment. The AI Hub corpus is composed of 1,602,708 sentences. The average syllable length and average word segment length for the Korean data were 58.58 and 13.53, respectively. The average syllable length and average word segment length for English were 142.70 and 23.57, respectively. The data from OpenSubtitles comprised a total of 929,621 sentences. The average syllable length and average word segment length for Korean were 21.82 and 5.5, respectively. The average syllable length and word segment length for English were 42.61 and 8.06, respectively.

To achieve subword tokenization, training data was pre-processed using SentencePiece [9] from Google, and the vocabulary size was set to 32,000. In the case of OpenSubtitles, data with fewer than three word tokens were deleted; this filtering process was not utilized for the AI Hub data. To form the validation set, 5000 sentences were randomly selected from the training data. The accuracy of every translation result was evaluated based on the BLEU score; this process used the multi-bleu.perl script of Moses (https://github.com/moses-smt/mosesdecoder/blob/master/scripts/generic/multi-bleu.perl). Table 1 presents an example of the training data.

Table 1. Training data example.

Source Sentence	Target Sentence
저는 당신의 컴퓨터 고치려고 이곳에 왔습니다.	I've come here to fix your computer.
이 이벤트를 위한 당신의 코디네이션에 감사드립니다.	Thanks for your coordination for this event.
당신의 행동은 오늘 나를 실망시켰어요.	Your actions disappointed me today.
당신은 직장에서 걱정할 게 정말 많네요.	You have lots of things at work to worry about.

The IWSLT-16 and IWSLT-17 datasets were used as the test set. This test set was established based on the TED domain; a substantial amount of Korean–English NMT research has been conducted based on this test set [32,33]. IWSLT-16 and IWSLT-17 are composed of 1143 and 1429 sentences, respectively.

4.2. Models and Hyperparameters

Because the primary aim in this study is to determine whether NMT model performance can be improved through various decoding strategies without changing the model, the model that was evaluated was based on the KU model, which was produced on the basis of the Transformer model [17], which is currently the most commercialized NMT model.

The hyperparameters of this model were set as follows: the batch size was 4096, 6 attention blocks and 8 attention heads were used, and the embedding size was 512. Adam and noam decay were used for optimization. Two GTX 1080 GPUs were utilized in the training process.

First, the performances of this model and previous Korean–English NMT models were compared. Specifically, the comparison was conducted with results from previous studies in which performances were evaluated with the same test set. The performance comparison included a study in which multilingual machine translation was explored to improve low-resource language translation [33] and a Korean–English NMT [32] that used MASS. The overall comparison results are listed in Table 2.

According to the comparison results, the KU model showed the best performance with IWSLT-16 scored 17.34, and the model that applied back translation to the MASS model of Kangwon National University showed the best performance with IWSLT-17 scored 15.34. It appears that the KU model showed the best performance owing to the quality of the AI Hub data. In addition, because the purpose of this model was to verify the quality of the data, data augmentation techniques such as back translation were not used. The model outperformed existing models when processing IWSLT-16 even though back translation was not applied; when processing IWSLT-17, the model appeared to outperform the existing models that did not employ back translation. It appears that the MASS model showed comparatively high performance because the IWSLT 2017 TED English–Korean Parallel Corpus was used as the learning data, and the test sets used in this study and the MASS study also utilized the same domain. In the case of the KU model, the data were not used as training data because it was judged unfair to use a domain that has the same test set as the training data. This is because if an NMT model that is specialized for a particular domain is deduced by some chance, it may become a structure in which the performance from the test set appears high, naturally.

4.3. Beam Size Comparison Experiment

To compare the performances of various decoding strategies, the performance impact of changes in beam size was measured by incrementing it from 1 to 10. The results of this experiment are shown in Table 3.

According to the results, the BLEU score could be a maximum of 1.31 depending on the change in the beam size. For both test sets, optimal performance was shown when the beam size was two. In other words, controlling the beam size directly affects the overall performance of the model, implying that decoding should be conducted using an optimal beam size to maximize model performance. In addition, it was shown that a large beam size does not necessarily guarantee high performance; in fact, the worst performance was observed when the beam size was 10, which is its largest setting. It appears that a large beam size negatively affects overall performance because it increases the number of candidates that must be calculated. An example of this appeared to occur when greedy decoding was conducted with a beam size of 1; this case showed that even a single incorrect prediction can fatally affect overall performance.

An experiment comparing decoding speeds according to beam size was also conducted. The most important factor in NMT processing is speed, and many organizations consider this factor when developing NMT. Problems such as server overload and outages can occur if slow speeds are encountered when the service is provided through a web server. Therefore, we determined that translation speed is the most important factor in NMT and conducted an experiment comparing the total translation time (Total Time) for the test set, average translation speed per sentence (Average), and number of tokens that can be processed per second (Token per/s) according to the beam size. The speed experiment was conducted using IWSLT-16. The results of the experiment are presented in Table 4.

Table 2. Performance comparison of existing studies on Korean–English neural machine translation (NMT) models.

Model	IWSLT-16	IWSLT-17
Out-domain Resources [33]	10.09	8.98
MASS [32]	15.57	13.50
MASS+BT [32]	17.11	**15.34**
Google Translate	15.28	13.03
KU [37]	**17.34**	14.75

Table 3. Experimental Results according to Beam size.

Beam Size	IWSLT-16	IWSLT-17
Beam 1	17.27	14.84
Beam 2	**17.77**	**15.19**
Beam 3	17.51	14.99
Beam 4	17.49	14.83
Beam 5	17.34	14.75
Beam 6	16.97	14.49
Beam 7	16.81	14.41
Beam 8	16.78	14.31
Beam 9	16.67	14.29
Beam 10	16.46	14.23

Table 4. Decoding speed according to beam size.

Beam Size	Total Time	Average	Token per /s
Beam 1	13.929	0.012	1609.359
Beam 2	14.667	0.012	1477.046
Beam 3	15.711	0.013	1353.141
Beam 4	16.241	0.014	1292.145
Beam 5	17.683	0.015	1175.981
Beam 6	18.565	0.016	1101.098
Beam 7	19.679	0.017	1026.473
Beam 8	20.949	0.018	960.227
Beam 9	22.693	0.019	881.692
Beam 10	23.907	0.020	828.938

According to the results, translation speed decreased as the beam size increased. Specifically, the translation speed for the entire test set, translation speed per sentence, and number of tokens processed per second all decreased proportionally as the beam size increased. According to Tables 3 and 4, a large beam size did not guarantee good performance and high speeds; rather, a moderate beam size increased both performance and speed. Figure 3 shows a graph of the relationship between the BLEU score and the translation speed for the entire test set.

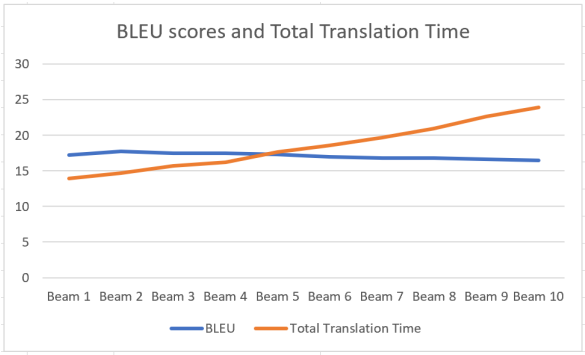

Figure 3. Graph of BLEU scores and translation speed.

As shown in Figure 3, speed decreases and performance degrades gradually as the beam size increases. In particular, the difference in speed between beam sizes 1 and 10 was 9.978 s of Total Time, which implies that organizations should carefully select the beam size for decoding if processing speed is a priority.

4.4. N-Gram Blocking Comparative Experiment

A second experiment was conducted to evaluate the performance achievable with n-gram blocking. In RBMT, sentence structures are rarely scattered, and words are rarely translated repetitively because translations are conducted according to stringent rules. However, in NMT, sentence structures are often scattered because issues such as repetitive translations and omissions occur owing to UNK problems caused by the limited vocabulary size. Copy Mechanism [39] was proposed to solve this problem; however, it requires complex changes to the model structure and decreases the processing speed. To solve this problem without changing the model structure, an experiment was conducted to prove that model performance could be improved by applying n-gram blocking during decoding. The results of the experiment are shown in Table 5. A beam size of two was selected for each model when n-gram blocking was utilized, reflecting the experimental results in Table 3.

According to the results of the experiment, performance stabilization was achieved with 8-gram blocking on both test sets, and performance was optimized when 8-gram and 4-gram blocking were applied on both test sets. In contrast, the performance decreased sharply when unigram blocking was applied, implying that this technique should not be applied in practice. In conclusion, performance improved compared to the performance achieved with IWSLT-17, as shown in Table 2. The results verify that the application of n-gram blocking led to a moderate improvement in performance.

4.5. Comparative Experiment on Coverage and Length Penalty

Finally, an experiment was conducted to determine if performance improves when SLP and ALP are applied. The results are shown in Table 6.

Table 5. Experimental results according to N-gram blocking.

N-gram Blocking	IWSLT-16	IWSLT-17
Uni-gram	5.14	4.98
Bi-gram	15.98	14.43
Tri-gram	17.62	15.09
4-gram	17.65	**15.24**
5-gram	17.74	15.22
6-gram	17.75	15.21
7-gram	17.72	15.20
8-gram	**17.77**	15.20
9-gram	**17.77**	15.20
10-gram	**17.77**	15.20

Table 6. Experimental results according to length penalty.

Penalty	IWSLT-16	IWSLT-17
Average Length Penalty	17.94	**15.42(+0.08)**
Stepwise Length Penalty	17.79	14.95
(Average+Step Wise) Length Penalty	**17.98(+0.64)**	15.22

According to the results, performance was optimized when SLP and ALP were applied for the processing of IWSLT-16; when only ALP was applied for the processing of IWSLT-17, not only was the best performance achieved but the BLEU score was also 0.08 points higher than the best BLEU score (15.34) reported in the existing Korean–English NMT research. When the length penalty was applied, the performance improvement was greater than when n-gram blocking was applied; this may have

occurred because these strategies mitigate the unfairness of the reduction in cumulative probability that occurs when a long sentence is translated.

We applied various decoding strategies, and optimal performance was shown during the processing of both IWSLT-16 and IWSLT-17, which are the most objective test sets in Korean–English NMT. In conclusion, it was confirmed that performance improvements could be achieved through various decoding strategies without changing the model. This methodology can aid low-resource NMT research in which it is difficult to obtain data, and this study should provide the basis for pre-processing and post-processing research in the NMT field.

5. Conclusions

Through various experiments, we proved that the performance of Korean–English NMT can be increased through various decoding strategies without changing the model structure. The performance was compared against those reported in previous Korean–English NMT studies, and IWSLT-16 and IWSLT-17 were used as test sets to maintain objectivity in performance evaluations. Experimental results obtained with a beam size of 2, n-gram blocking, and a length penalty showed performance that was comparatively better than those reported in previous Korean–English NMT studies. Additionally, the model was distributed as a platform to increase its availability. By using various decoding strategies that were proven in this study, both speed and performance were improved without changes to the model structure. Performance may be further improved by integrating ideas proposed in other studies in which models were lightened through strategies such as network pruning [40], knowledge distillation [41], and quantization [15]. Hence, an effective beam search technique and a new decoding technique will be investigated in future studies. In addition, an optimal NMT service system will be researched by applying the various decoding strategies in conjunction with network pruning, knowledge distillation, and quantization. A limitation of this study is the lack of an innovative new decoding strategy. Therefore, in the future, we plan to investigate decoding strategies that are more efficient.

Author Contributions: Funding acquisition, H.L.; investigation, C.P.; methodology, C.P.; project administration H.L.; conceptualization, C.P. ; software, C.P.; validation, C.P.; formal analysis, C.P.; investigation, C.P.; writing—review and editing, Y.Y., K.P., H.L. and H.L.; supervision, H.L.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute for Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government Ministry of Science and ICT (MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques), and this research was supported by the MSIT, Korea, under the ITRC (Information Technology Research Center) support program (IITP-2020-2018-0-01405) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation).

Acknowledgments: Thanks to AI Hub for creating a great dataset. I am very grateful to my friend Yeonsu Lee (Sungkyunkwan University) for helping me with English corrections.

Conflicts of Interest: The authors declare no conflict of interest

References

1. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
2. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
3. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2019; pp. 5754–5764.
4. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
5. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* **2019**, arXiv:1910.10683.

6. Park, C.; Kim, G.; Lim, H. Parallel Corpus Filtering and Korean-Optimized Subword Tokenization for Machine Translation. In Proceedings of the 31st Annual Conference on Human & Cognitive Language Technology, Busan, Korea, 11–12 October 2019; pp. 221–224.
7. Park, C.; Kim, K.; Lim, H. Optimization of Data Augmentation Techniques in Neural Machine Translation. In Proceedings of the 31st Annual Conference on Human & Cognitive Language Technology, Busan, Korea, 11–12 October 2019; pp. 258–261.
8. Sennrich, R.; Haddow, B.; Birch, A. Neural machine translation of rare words with subword units. *arXiv* **2015**, arXiv:1508.07909.
9. Kudo, T.; Richardson, J. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv* **2018**, arXiv:1808.06226.
10. Koehn, P.; Guzmán, F.; Chaudhary, V.; Pino, J. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), Florence, Italy, 1–2 August 2019; pp. 54–72.
11. Chen, W.; Matusov, E.; Khadivi, S.; Peter, J.T. Guided alignment training for topic-aware neural machine translation. *arXiv* **2016**, arXiv:1607.01628.
12. Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding back-translation at scale. *arXiv* **2018**, arXiv:1808.09381.
13. Currey, A.; Miceli-Barone, A.V.; Heafield, K. Copied monolingual data improves low-resource neural machine translation. In Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, 7–11 September 2017; pp. 148–156.
14. Chatterjee, R.; Federmann, C.; Negri, M.; Turchi, M. Findings of the WMT 2019 Shared Task on Automatic Post-Editing. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), Florence, Italy, 1–2 August 2019; pp. 11–28.
15. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
16. Ng, N.; Yee, K.; Baevski, A.; Ott, M.; Auli, M.; Edunov, S. Facebook FAIR's WMT19 News Translation Task Submission. *arXiv* **2019**, arXiv:1907.06616.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2017; pp. 5998–6008.
18. Cettolo, M.; Jan, N.; Sebastian, S.; Bentivogli, L.; Cattoni, R.; Federico, M. The IWSLT 2016 evaluation campaign. In Proceedings of the International Workshop on Spoken Language Translation, Seattle, WA, USA, 8–9 December 2016.
19. Cettolo, M.; Federico, M.; Bentivogli, L.; Jan, N.; Sebastian, S.; Katsuitho, S.; Koichiro, Y.; Christian, F. Overview of the iwslt 2017 evaluation campaign. In Proceedings of the 14th International Workshop on Spoken Language Translation Tokyo, Japan, 14–15 December 2017; pp. 2–14.
20. Kasher, A. *Language in Focus: Foundations, Methods and Systems: Essays in Memory of Yehoshua Bar-Hillel*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012; Volume 43.
21. Dugast, L.; Senellart, J.; Koehn, P. Statistical Post-Editing on SYSTRAN's Rule-Based Translation System. In Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, 23 June 2007; pp. 220–223.
22. Forcada, M.L.; Ginestí-Rosell, M.; Nordfalk, J.; O'Regan, J.; Ortiz-Rojas, S.; Pérez-Ortiz, J.A.; Sánchez-Martínez, F.; Ramírez-Sánchez, G.; Tyers, F.M. Apertium: A free/open-source platform for rule-based machine translation. *Mach. Transl.* **2011**, *25*, 127–144. [CrossRef]
23. Zens, R.; Och, F.J.; Ney, H. Phrase-based statistical machine translation. In *Annual Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 18–32.
24. Koehn, P. *Statistical Machine Translation*; Cambridge University Press: Cambridge, UK, 2009.
25. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
26. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.

27. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional sequence to sequence learning. In Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1243–1252.
28. Wu, F.; Fan, A.; Baevski, A.; Dauphin, Y.N.; Auli, M. Pay less attention with lightweight and dynamic convolutions. *arXiv* **2019**, arXiv:1901.10430.
29. Lample, G.; Conneau, A. Cross-lingual language model pretraining. *arXiv* **2019**, arXiv:1901.07291.
30. Song, K.; Tan, X.; Qin, T.; Lu, J.; Liu, T.Y. Mass: Masked sequence to sequence pre-training for language generation. *arXiv* **2019**, arXiv:1905.02450.
31. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *arXiv* **2020**, arXiv:2001.08210.
32. Jeong, Y.; Park, C.; Lee, C.; Kim, J.s. English-Korean Neural Machine Translation using MASS. In Proceedings of the 31st Annual Conference on Human & Cognitive Language Technology, Busan, Korea, 11–12 October 2019; pp. 236–238.
33. Xu, G.; Ko, Y.; Seo, J. Low-resource Machine Translation by utilizing Multilingual, Out-domain Resources. *J. KIISE* **2019**, *40*, 649–651.
34. Lee, J.; Kim, B.; Xu, G.; Ko, Y.; Seo, J. English-Korean Neural Machine Translation using Subword Units. *J. KIISE* **2018**, 586–588.
35. Xu, G.; Ko, Y.; Seo, J. Expanding Korean/English Parallel Corpora using Back-translation for Neural Machine Translation. In Proceedings of the 30th Annual Conference on Human and & Cognitive Language Technology, Seoul, Korea, 12–13 October 2018; pp. 470–473.
36. 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, 7–12 July 2002, pp. 311–318.
37. Park, C.; Lim, H. A Study on the Performance Improvement of Machine Translation Using Public Korean-English Parallel Corpus. *J. Digit. Converg.* **2020**, *18*, 271–277.
38. Provilkov, I.; Emelianenko, D.; Voita, E. BPE-Dropout: Simple and Effective Subword Regularization. *arXiv* **2019**, arXiv:1910.13267.
39. Gu, J.; Lu, Z.; Li, H.; Li, V.O. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv* **2016**, arXiv:1603.06393.
40. Molchanov, P.; Mallya, A.; Tyree, S.; Frosio, I.; Kautz, J. Importance estimation for neural network pruning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11264–11272.
41. Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.