Smart Media Journal / Vol.12, No.1 / ISSN:2287-1322
https://dx.doi.org/10.30693/SMJ.2023.12.1.32

Korean Text to Gloss: Self-Supervised Learning approach

Thanh-Vu Dang, Gwang-hyun Yu, Ji-yong Kim, Young-hwan Park, Chil-woo Lee, Jin-Young Kim

Abstract

Natural Language Processing (NLP) has grown tremendously in recent years. Typically, bilingual, and multilingual translation models have been deployed widely in machine translation and gained vast attention from the research community. On the contrary, few studies have focused on translating between spoken and sign languages, especially non-English languages. Prior works on Sign Language Translation (SLT) have shown that a mid-level sign gloss representation enhances translation performance. Therefore, this study presents a new large-scale Korean sign language dataset, the Museum-Commentary Korean Sign Gloss (MCKSG) dataset, including 3828 pairs of Korean sentences and their corresponding sign glosses used in Museum-Commentary contexts. In addition, we propose a translation framework based on self-supervised learning, where the pretext task is a text-to-text from a Korean sentence to its back-translation versions, then the pre-trained network will be fine-tuned on the MCKSG dataset. Using self-supervised learning help to overcome the drawback of a shortage of sign language data. Through experimental results, our proposed model outperforms a baseline BERT model by 6.22%.

Keywords: Sign Language Production | Neural Machine Translation | Korean Corpus.

I. INTRODUCTION

As a primary communication means of the deaf community, sign languages have their own grammatical rules and linguistic structures yet are scarcely known by the rest. Sign language is the visual language utilizing multiple channels to convey information. These features are varied, such as hand and mouth shape, facial expression, and head, shoulders and torso movement.

Currently, state-of-the-art sign interpreters use gloss-level tokens in intermediate translation, where glosses are minimal lexical entities representing singular signs. Using sign glosses is vital because sign language grammar does not share with their spoken counterpart. The differences can be varied, such as word order, multiple channels used to convey concurrent information, and temporal/spatial details to describe the relationships between objects. The mapping between speech and sign is sophisticated, and there is no word-to-sign mapping.

Like spoken languages, sign language also varies across regions and cultures. Compared to widely used languages like English and Chinese, Korean is often referred to as a low-resource language in the research community. The situation is even worse for the Korean sign language as it is far more difficult to collect high-quality data[1]. Most existing sign language datasets are limited to a small number of words. Due to the inadequate vocabulary size, models learned from those datasets could be applied in practice.

* This research is supported by the Ministry of Culture, Sports, and Tourism (MCST) and 373 Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & 374 Development Program (R2020060002).

Manuscript : 2022. 11. 30
Revised : 2022. 12. 23, 2023. 02. 02
Confirmation of Publication: 2022. 02. 09
Corresponding Author : Jin-Young Kim e-mail : beyondi@jnu.ac.kr

This paper is motivated to provide assistive technology that enables deaf people to communicate in their language. We aim to contribute a new dataset for linguistic research about Korean Sign Language. Specifically, our main contributions are summarized as follows:

- We collect the sign language dataset containing 3828 pairs of Korean sentences and their corresponding sign glosses. We name the dataset as Museum-Commentary Korean Sign Gloss (MCKSG) dataset. The dataset is restricted to the museum context to describe historical merchandise and museum announcements. The dataset is provided with careful revision from Korean sign language experts.
- We present a framework for translation from Korean text to Sign Gloss. The framework is two folds based on self-supervised learning. The pretext task is a text-to-text from a Korean sentence to its back-translation versions from English; then, the pre-trained network will be fine-tuned on the MCKSG dataset, where a Korean input sentence will be translated to its corresponding sign glosses.

II. RELATED WORK

1. *Sign Language Dataset*

One of the challenges in SLT research is the availability of sufficient datasets. There are two reasons why the sign language dataset has yet to be assembled thoroughly. Firstly, sign language is mainly used by the deaf community, a minority in the predominant auditory world. Like spoken languages, moreover; sign languages also vary among regions and communities. Secondly, curating and annotating continuous sign language videos with spoken subtitles is non-trivial. The available annotations need to be more comprehensive to build models that work on a large domain, either in quality or quantity. Both these peculiarities and deficiencies prevent the development of sign language research.

The first study that introduced large-scale sign language datasets was from Camgoz et al.[2]. They presented the PHOENIX14T dataset, which provides spoken language translations and gloss-level annotations for German sign language videos of weather broadcasts. The dataset contains over .95M frames from a sign vocabulary of 1K signs and a German vocabulary of 2.8K words. About English sign language, Li et al.[3] constructed a large-scale signer-independent WLASL video dataset, which contains over 68K videos of over 2000 signs performed by more than 100 signers from multiple educational sign language websites.

For Turkish, Ozge et al.[4] presented a Turkish sign language dataset containing over 38k sign videos of 226 signs performed by 43 signers. For the Korean language, no other research except for the work of Ko et al.[5] introduced the very first systematically large-scale Korean sign language dataset, namely KETI. The KETI dataset was collected from 14 signer experts who are deaf or hard of hearing. The dataset exposes various emergency contexts extracted from 14,672 high-resolution videos that recorded the Korean signs corresponding to 419 words and 105 sentences.

2. *Sign Language Recognition and Translation*

Earlier studies have considered SLR a gesture recognition problem. SLR seeks to recognize a continuous sequence of signs but ignores sign language's linguistic

properties and grammatical structures. Rastgoo et al. [6] proposed a pipeline consisting of a single-shot detector, 2d/3d CNN and LSTM to estimate 3D hand key points and recognize gestures from that 3D hand skeletons. Several other studies have focused on developing aid-device with built-in hand gesture recognition functions. Those systems' input might vary depending on the device, such as signal data obtained from sensors of hand gloves [7, 8].

In contrast, Sign Language Translation aims to generate continuously spoken language translations from sign language videos. Camgoz et al. [2] formalized the SLT framework as a neural machine translation task with the seq2seq model to learn the Spatiotemporal representation of the signs and spoken language. Several studies on video-to-text SLT have been conducted and benchmarked on the PHOENIX14T dataset. Typically, the research group of Camgoz published a new benchmark on their PHOENIX14T dataset, which has surpassed their previous results by replacing the seq2seq back-bone model with Transformer [9]. Zhou et al. introduced spatial-temporal multi-cue (STMC) [10] into the deep sequential model to preserve the uniqueness and explore the synergy of different cues both in terms of spatial and temporal information. Research from Yin et al. [11] reported state-of-the-art results on the PHOENIX14T dataset by interjecting the STMC module into Transformer.

3. *Sign Language Production*

Generating sign language from spoken language is a complex task that cannot be accomplished with trivial one-to-one correspondence. In addition, the number and order of glosses do not match the words of the spoken language sentence, so context and meaning conveyed from spoken to sign language are generally lost.

Instead of directly generating visual content, SLP can be divided into sub-tasks, as shown in Figure 1 (right), glosses sequence generating, and photo-realistic sign language mapping [12]. This study concentrates on the early sub-process to treat SLP as a translation from spoken to sign language problem. From the obtained sequence of sign glosses, sign languages are performed by avatars manipulating or realistic signer videos.

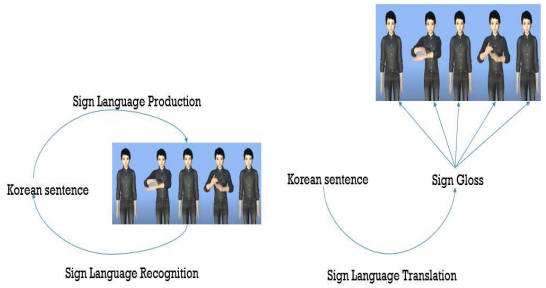


Fig. 1. Left: Sign Language Translation and Production. Right: Sign Language Translation with gloss as intermediate output.

Compared to sign language recognition and translation, only a few studies have progressed in sign language production [6, 8]. Stephanie et al. [12] combined NMT. with a motion graph to translate spoken language into sign poses and produced sign videos by generative model conditioned on resulting poses. A progressive Transformer with novel counter positional embedding [13] is introduced by Ben et al. to generate continuous 3D sign pose sequences. To assist non-signers in performing sign language, Ventura et al. [14] applied motion transfer [15] for generating synthesis sign videos given pose key points extracted from OpenPose [17]. We summarize public sign language datasets with their benchmarks in Table. 1.

Table 1. Public sign language dataset and evaluations.

Dataset	Task	Method	Results
KETI 14,672 videos 419 words 105 sentences	Sign2Text	Seq2Seq (GRU) / Seq2SeqAtt / Transformer [5]	BLEU-1: 50.80/ 65.26/ 66.58 ROUGE-L: 90.03/ 96.63/ 94.14
PHOENIX14T (2018) 95K frames 1K signs 2.8K words	Gloss2Text	Transformer [9]/ Seq2SeqAtt (GRU) [2]/ STMC + Transformer [11]	BLEU-1: 48.90/ 44.13/ 48.40
	Sign2Gloss2Text		BLEU-1: 48.47/ 43.29/ 50.63
	Sign2Gloss -> Gloss2Text		BLEU-1: 47.74/ 41.54/ 47.49
	Sign2Text		BLEU-1: 45.34/ 32.24
	Text2Gloss	Symbolic + Progressive Transformer GAN Motion Graph [12]	BLEU-1: 55.18 / 50.67
	Text2Pose		BLEU-1: 31.36/ 12.38
	Text2Gloss2Pose		BLEU-1: 31.07
WLASL/ASL 68K videos 2K signs	Sign2Text	GRU / I3D / TGCN [3]	Top 1: 22.54 / 23.65 / 32.48 Top 10: 61.38/ 62.24/ 66.31
	Gloss2Text	STMC + Transformer [11]	BLUE-1: 92.88

Korean corpora still need to be supplied with great resources of text datasets that facilitate the NLP task. Won et al. [1] investigated 32 publicly available Korean corpora covering a wide range of research, yet not including any Sign language dataset. The lack of a big-scale dataset leads to the absence of a pre-trained language model for Korean and causes impediments to transfer learning. Although communication is two ways of interaction, previous studies have almost exclusively focused on Sign Language Translation, and

only a few works in literature have demonstrated Sign Language Production. Moreover, although previous research has presented many architectures, studies have yet to examine unsupervised learning for small-size sign language datasets, which is a non-trivial problem.

III. METHOD

1. *Museum-Commentary Korean Sign Gloss Dataset (MCKSG)*

Sign languages differ from hand languages, where the shape of a single hand represents each letter in an alphabet, while the linguistic meaning of each sign is determined by the subtle difference of shape and movement of the body, hands, and even by the facial expression of the signer. For the ease of representing sign language in writing, sign sentences can be shattered into a sequence of glosses, the lexical instances representing singular signs. Similar to spoken language, there is more than one combination of glosses to describe a single intention depending on the signer's wording. However, in general, the number of sign texts is far less than that of spoken dictionaries.

1.1. *Data Curation*

The process of data curation is divided into four main steps: reference collection, sign interpretation, data evaluation, and data formation.

Reference collection: The MCKSG dataset targets 3828 sentences of commentary and guide materials for the exhibition facilities of the Gwangju Museum. A sign language interpreter (expert in sign language, signer) produces reference materials for museum commentary texts collected jointly with the Seodaemun City Hall for the Hearing Impaired in Korea so that the deaf can understand them together.

Sign Interpretation: When filming the museum commentary for a signer, the focus was on conveying meaning by accurately applying Korean sign language grammar. The filmed museum commentary video is inspected by the hearing impaired to measure sentence comprehension and response accuracy, and the video that needs correction is modified through consultation with the hearing impaired based on the reference material for the museum commentary sentence.

Data evaluation: To accurately convey the meaning of sign language when editing images through consultation with the hearing impaired, icon words, blank spaces, number shapes, sign language, and sign language numbers that make sign language words longer than dictionaries were defined and applied. In addition, for the hearing impaired to feel the museum commentary more naturally, we defined and applied anhydrous symbols that can express the head, body, face, mouth, and movement of the face and body, sign language characters and variations, and language expressions of the hearing-impaired recognition symbol. Based on the final reviewed museum commentary video, we created a Korean sign language Gloss dataset for museum commentary (MCKSG).

Data formation: We analyzed the dataset to determine the frequency of words and sorted them in ascending order. Words such as glosses, proper nouns, and Chinese characters used more than three times were selected, duplicate words, synonyms, and homonyms were also reviewed to create a lexicon of the museum commentary Korean sign language gloss dataset. Lastly, the museum commentary Korean sign language lexicon is divided into six

categories: Sign Words, Similar Words, Iconography, Fingerspelling, Fingerspelling Numbers, and Non-resin symbols. The complete process of data curation is explained in Figure 2.

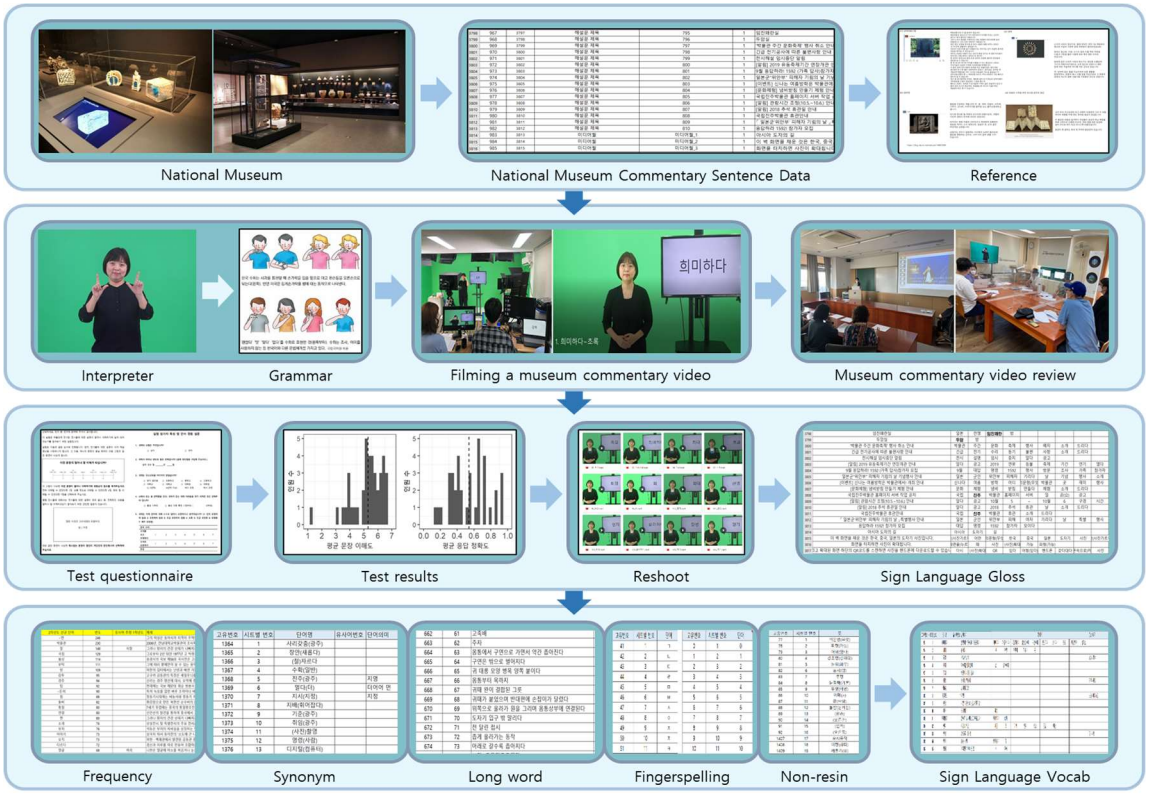


Fig. 2. Data Curation Process

1.2. *Vocabulary*

Sign Language Translation could be defined as a machine translation task, particularly in this study; the source language is Korean, and the target language is Sign Gloss. As the Sign glosses are notated by Hangul (Korean alphabet), the source and target vocabularies can be shared to unify identical tokens in both languages. Our sign vocabulary contains 3430 gloss tokens, covering a wide range of information, including number letters to describe the numerical quantity and Romanians to describe foreign characters. In detail, our sign vocabulary is constructed from six types of tokens, Sign Words, Similar Words, Iconography, Fingerspelling, Fingerspelling Numbers, and Non-resin symbols, which Korean Sign Language experts conceptualize. The gloss vocabulary is built mainly on analyzing the movement of an individual sign gloss so that each token in the Gloss vocab corresponds to an actual motion in Korean sign language.

Korean sign language experts have tokenized the target language, but the source language requires us to define another tokenizer for constructing its vocabulary. This study uses a Konlpy-based Open-Korean-text tokenizer [18] with lemmatization as a tokenizer for the Korean source sentence. We further replace tokens related to numerical information with a special token,

"Number," and replace Romanian tokens with another special token, "Alpha". This replacement improves our vocabulary compactness since those tokens rarely appear a second time, although each conveys consistent information for an individual sentence. "Number" and "Alpha" tokens represent co-existent information in the source sentence and target glosses. We note that a simple post-preprocessing can help recover those tokens' original content. Table 2 shows examples in the MCKSG dataset and corresponding tokens. Using the KoNLPy tokenizer with our modification on "Number" and "Alpha" tokens provides a vocabulary size of 7234 that covers all source sentences in the MCKSG dataset. We show quantitative information on the MCKSG dataset in Table 3.

Table 2. Examples from the MCKSG dataset

Input	Tokens	Target (Glosses)
한판 전남 동부지역은 4 세기 후반~5 세기 전반부터 6 세기 전반까지 가야계문화가 확인된다.	'한편', '전남', '동', '부지역', 'Number/4 세기', '후반', 'Number/5 세기', '전반', 'Number/6 세기', '전반', '가다', '문화', '확인', '되다'	'전남', '동', '지역', 'Number/4C', 'Number/5C', '전', '부터', 'Number/6C', '전', '가야', '관계', '문화', '확인', '완료', '휴지동작', '휴지(쉼)
몸의 형태에 따라 버들잎 모양과 비파형으로 구분된다.	'몸', '형태', '따르다', '버들잎', '모양', '비파', '형', '구분', '되다'	''몸', '모습', '따르다', '잎', '모습', '또한', '비파', '모습', '이층삼각형', '나누다', '따로'
그리고 확대된 화면 하단의 QR 코드를 스캔하면 사진을 핸드폰에 다운로드할 수 있습니다.	'그리고', '확대', '되다', '화면', '하단', 'Alpha/QR', '코드', '스캔', '사진', '핸드폰', '다운로드', '수', '있다'	'다시', ' (사진)확대', ' QR', ' 있다', ' 어형(있다)', ' 핸드폰', ' 갖다대다', ' (핸드폰속으로)저장되다', ' 사진', ' 다운받다', ' 가능', ' 파형(가능)'

Table 3. A vocabulary of the MCKSG dataset

	Token's types	Number of tokens	Examples
Source Vocabulary	Korean words	7234	떨어지다, 문화, Alpha, Number
Target Vocabulary	Sign Words	1207	기도, 떨어지다
	Similar Words	163	디지털(컴퓨터)
	Iconography	436	v(체크), V(불꽃무늬)
	Fingerspelling	41	ㄱ, ㅐ
	Fingerspelling Number	130	0, 10, 100
	Non-resin symbols	19	의문형(무엇), 강조형(강하다)

2. *Framework*

Generating sign language videos from given spoken sentences is a complicated task due to the long-term dependency on sequential context could not be retained in visual context by SLP architecture. Because of this situation, we treat Sign Language Production (SLP) as two sub-processes: a translation task from spoken language into sign glosses and a one-by-one mapping between a gloss and a visual motion. This study focuses on the early step because the sign gloss alone is accurate enough to represent visual matter for signs, as shown in Figure 1.

To obtain the text-2-gloss model, we propose a two-phase process based on self-supervised learning, including pretext and downstream tasks. Figure 3 explains our overall framework. The input of our model is a sentence written in

Korean $\boldsymbol{x} = x_1 x_2 \dots x_n$, where x_i is a word or token after tokenizing. Our model aims to generate a sequence of sign glosses $\boldsymbol{g} = g_1 g_2 \dots g_n$, where $g_i \in \mathbb{G}$ is a gloss identified in our gloss vocabulary G, as introduced in Section II.1. The objective model f is fine-tuned on the MCKSG dataset for a downstream task as machine translation, while the pre-trained model f_{pt} is obtained by training on a larger and more general dataset.

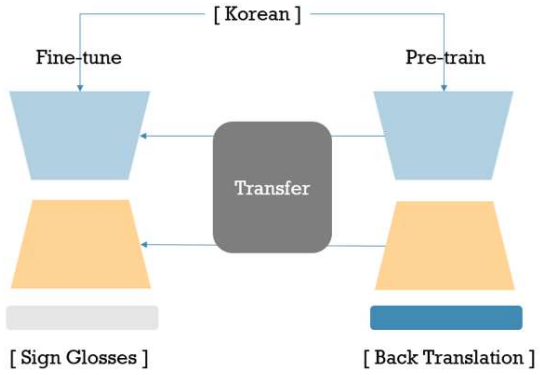


Fig. 3. Proposed framework: the pretext task is defined as machine translation from a Korean sentence to its back translation, and the downstream task is also a machine translation task but from a Korean sentence to its corresponding sequence of sign gloss. The pre-trained model from the pretext task will be fully transferred to the downstream task except for the final layer of the decoder.

2.1. *Tokenizer*

Tokenizing a sentence is splitting it into words or subwords, identified in a vocabulary provided by an index in a look-up table. For other languages different from English, sophisticated rules are added into the tokenizer for accurately splitting a sentence into components that are not necessarily a word in a dictionary but either a word, subword or compound word. Subword tokenizer is a hybrid between word-level and character-level tokenizer that [18] has been employed extensively in Transformers [19, 20, 21] for their effectiveness. This study mainly experiments with WordPiece [22] tokenizer seeded on pre-tokenization using the KoNLPy package.

Table 4. Tokenization

Input	WordPiece Tokens
토기 아가리 모양이 다양하며, 작은 그릇도 만들었다.	'[CLS]', '토', '##기', '아가', '##리', '모양', '다양', '##하며', '작은', '그릇', '만들었다', '[SEP]'
삼국시대에 전남지역에는 주변지역을 지키고 통제하기 위한 성들이 많이 만들어졌다.	'[CLS]', '삼국시대', '전남', '지역', '주변', '지역', '지키', '##고', '통제', '하기', '위', '[SEP]'
국립중앙박물관 불교회화실에 높이 12m 폭 6m 의 괘불 미디어아트가 펼쳐집니다.	'[CLS]', '국립', '중앙', '박물관', '불교', '회화', '실', '높이', 'Number', 'Alpha', '폭', 'Number', 'Alpha', '의', '괘', '##불', '미디어', '##아', '##트', '펼쳐', '##집니다', '[SEP]'

The sign (target) sentences have already been tokenized into a sequence of glosses, yet our Korean input sentence (source) needs to be tokenized for further processes. In this study, we use Konlpy-based Open-Korean-text as our tokenizer on account of a specialized tool for the Korean language. Additionally, we modify the tokenization slightly to unify Romanian words and numbers. We replace each token with a special token "Number," to indicate that the token is a number and "Alpha" to indicate that the token is a Romanian word or keep it unchanged depending on its parts of speech. Our tokenizer follows the pipelines of 4 steps, normalization, pre-tokenization, model and post-processing. For normalization, we apply standard preprocessing techniques such as stripping white spaces, lowering all characters and removing punctuation and Korean stop words. For

pre-tokenization, we prepare a set of tokens using the Konlpy tokenizer with the special "Number" and "Alpha" tokens. We train a WordPiece model founded on the pre-tokens set. The role of the WordPiece model is to further split a word into tokens and is responsible for mapping those tokens to their corresponding IDs in the vocabulary. Finally, we construct a template for a sentence with a "[BOS]" token at the head and a "[EOS]" token at the end of that sentence; worth to notice that we also add a "[PAD]" token when processing on a batch of sentences. Table 4 shows examples of output from the tokenization step. We utilize Wordpiece tokenization to build a united vocabulary for pretext and downstream tasks. In particular, our vocabulary is constructed from 1.8M sentences of 6 Korean datasets for various tasks. Using the WordPiece model, our vocabulary size is set as 10K tokens instead of over 80k tokens at the word level. However, to unite the vocabulary for downstream task usable, we need to expand the vocabulary to involve gloss tokens. Our vocabulary size is 12078, including tokens from the extra datasets, MCKSG gloss and special tokens.

2.2. *Pretext task: Back Translation*

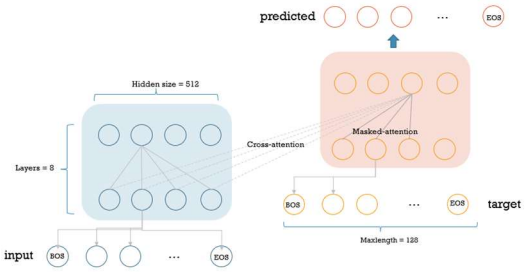


Fig. 4. Overall architecture. Our model for the pretext task is an encoder-decoder, Bert-based model. The input of the encoder is a Korean sentence with a maximum length is 128, including that sentence's word tokens, [BOS], [EOS] and [PAD] tokens. The output of the encoder is embeddings of tokens at every layer, which will be fetched into the decoder's hidden layers to calculate cross-attention scores. The decoder is a language model that generates a sequence of tokens which later will be matched to the true label. The input of the decoder is the corresponding back translation of the source sentence with the same format. While the self-attention mechanism in the encoder is bidirectional, the decoder employs masked attention to avoid leaking future information.

The main issue of the Sign Language Translation task needs labeled data since it is a domain-specific task. A frequently used strategy to address this problem is to transfer the knowledge from a model that has already been trained on a large dataset to our domain. However, the task of interest in this study meets another challenge: the dataset is given in the Korean language. There are few pre-trained models available, and there is no pre-trained language model for Korean [23, 24]. For proper knowledge transfer from a pre-trained model, this study designs a pretext task with a decoder head that is straightforward to transfer to the downstream Sign Language Translation task. Analysis of the MCKSG dataset reveals that the target sequence of sign glosses shares a major token with the input sentence. We found that the overlapping portion is 41% on average. Based on this observation, we design a pretext task that resembles the downstream job but emphasizes two different aspects. First, the pretext task should be a generator that can generate a sequence of tokens. Secondly, the source and target can share the same vocabulary. Back translation can be employed as a pretext task because it satisfies both conditions and works as a paradigm like sign language translation, the downstream task of interest.

Self-supervised learning can overcome the shortage of labeled data by introducing a pre-trained pretext task on a large unlabeled or pseudo-labeled dataset. We trained. The model and tokenization for the pretext task in this study are on four Korean datasets. We built the dataset for the pretext task by translating a source sentence from Korean to English and back to Korea. The back translation [25] is considered an unsupervised data augmentation method [25, 26]. The counterpart back translation from any language can be effortlessly obtained from the input using state-of-the-art machine translators. Back translation also works as normalization, where the translated sample is paraphrased in a pre-defined structure set by a translator. To avoid stylistic bias in generating data. Therefore, we utilized several translators to synthesize translated samples. The tools used in this study include Google Translator API [26], Papago Translator API and EasyNMT – a pre-trained machine translation that supports over 150 languages [27]. In this study, the pretext is modeled as an encoder-decoder, where both the encoder and decoder are based on Bert's [20] architecture. The overall architecture is shown in Figure 4.

2.3. *Downstream task: Sign Language Translation*

Our downstream task of interest is Sign language translation which takes a Korean sentence as input and produces a sequence of sign glosses as output. The pretext task mentioned above has been designed so that it is ready to fully transfer knowledge to the downstream task since the downstream task can reuse the vocabulary from the pretext task, and the encoder-decoder architecture is also appropriate to constitute the learning model for the downstream task. The only difference between the downstream task and the pretext task is how to set the tokenizer for the target language; while the encoder receives tokens that the above scheme has tokenized, we map the glosses of sign sentences into pre-defined IDs in our vocabulary.

IV. Experimental results

1. *Settings*

Dataset: We trained the pretext task on a back translation dataset synthesized from four extra Korean datasets and the Korean input sentences from the MCKSG dataset. The details of these datasets [30] are given as follows:

•KorNLI: the dataset is built for the natural language inference task, containing 950,354 pairs of sentences. The training set size is 942,854, the development set is 2,490, and the test set size is 5,010. After performing back translation and selection, we obtain 80,0410 pairs of sentences.

•Naver-ner: the dataset is built for name entity recognition, containing 90,000 sentences. The training set has a size of 81000, and the test set size is 9,000. After performing back translation and selection, we obtain 88,000 pairs of sentences.

•korSTS: the dataset is built for the semantic textual similarity task, containing 8,628 pairs of sentences. The training set size is 5,749, the development set is 1,500, and the test set is 1,379. After performing back translation and selection, we obtain 17,262 pairs of sentences.

•Question-pair: the dataset is built for the duplication-checking task, comprising 15,512 sentences. The training set size is 6,136 pairs, the development set is 682 pairs, and the test set is 758 pairs. After performing back translation and selection, we obtain 15,158 pairs of sentences.

•MCKSG: the dataset is built by ourselves

for sign language translation in a museum context, consisting of 3,828 pairs. We only performed back translation with Korean inputs and obtained 3,828 pairs of original Korean and their back-translated sentences.

Entirely we generate 924,658 Korean back translation sentences. The pretext model and tokenizer are trained and evaluated on this synthesis dataset, while the downstream model is only trained and evaluated on the MCKSG dataset. After training the WordPiece tokenizer, our vocabulary has a size of 10,000. We then add sign glosses and special tokens into our vocabulary to ensure they cover all the downstream task tokens. In total, our vocabulary has a size of 12,078.

Training:

Pretext task: The pretext model is an encoder-decoder Bert-based type where both the encoder and decoder share the same hyper-parameters: The number of layers is 8, the number of attention heads is 8, and the hidden size of each layer is 512. Self-attention in the encoder is bidirectional, but it is masked for the right-side nodes in the decoder to avoid leakage of future information when generating output. Besides, the decoder and encoder are connected by cross-attention. For this architecture, our model possesses 94.9M parameters.

We trained the pretext model for the back-translation task with the cross-entropy loss using AdamW optimizer, warming up learning at 0.0001. With a batch size of 32, the model takes 40 hours to train for ten epochs on double TITAN V GPUs. We split the back translation synthesis dataset into a training set and validation set with a ratio of 7:3. The best model was evaluated by cross-validation on the loss value of the validation dataset.

Downstream task: As mentioned above, the pretext model is ready for fully transferring to the downstream task since they both share the same vocabulary and architecture. We split the MCKSG dataset into the training and validation set with a ratio of 7:3. Since our dataset is relatively small and has high randomness, further splitting data for the test set will cause a lack of training data. For all experiments, we train the model with a batch size of 32 and SGD optimizer with a learning rate of 0.0001.

2. *Results*

According to sign language experts and evaluation with the hearing-impair person, the context of a Korean sign sentence needs to be stronger dependent on gloss order [5]. Therefore, this study suggests using the Jaccard score (JC) as an evaluation metric for sign language translation. In the data creation process, Korean signers varied the order of the ground-truth glosses, causing evaluation metrics such as BELU or ROUGE cannot truly reflect the model's performance. The JC score used in this study is inspired by calculating the overlapping portion of 2 bounding boxes in the object detection task. The equation of the JC score is given below.

$$JC(s_1, s_2) = \frac{|s_1 \cap s_2|}{|s_1 \cup s_2|}$$

Where s_1, s_2 is the set of tokens in sentence 1 and sentence 2, the JC score between two sentences is given by the cardinal of intersection over the union set. Note that the model with a higher JC score is better. For fair evaluation, we also performed cross-validation, and the results were reported in an average JC over ten runs.

Table 5 shows the performance of the downstream model on the CMSKG dataset. For comparison, we introduced two other

models trained on the small vocabulary of the MCKSG dataset with the size of 8K tokens at the word level. After 100 epochs of training or fine-tuning, the RNN with attention model achieve a JC score of 67.56, and the BERT-based small vocabulary without transfer achieves a JC score of 70.99, which is relatively lower than that of the BERT-based model on large vocabulary without transfer, 73.86, and transfer 77.21, respectively.

Table 5. Performance of four models (RNNA, BERT-based small/large vocabulary) on downstream task with the MCKSG dataset

Model	JC score (10 epochs)	JC score (100 epochs)
RNNA (small vocab)	–	67.56
BERT no transfer (small vocab)	–	70.99
BERT no transfer (large vocab)	36.43	73.86
BERT Transfer (large vocab)	76.62	77.21

Transfer learning usually obtains high performance within a smaller number of epochs than a model trained from scratch. To support the above account, we fine-tuned the downstream model for 10 and 100 epochs to observe how the transferred model behaves in the initial learning stage. As shown in Figure 5, the loss from the transfer model reduces faster than the one without transfer, proving that the pretext task transfers proper knowledge to the downstream task. Additionally, the training scheme shows that downstream models quickly overfit to the training dataset after 5k steps. Hence, we used early stopping to obtain the final model.

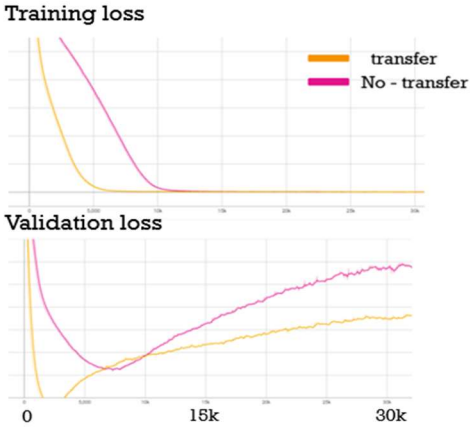


Fig. 5. Training (upper) and validation loss (below) on downstream tasks with and without transfer learning of BERT-based large vocab model.

V. CONCLUSIONS

This study considers a specific context domain, museum-commentary Korean sign language, and proposes a learning framework based on self-supervised learning that addresses the small dataset to assist the deaf community. The result ties well with previous studies wherein self-supervised learning shows superior performance on a small dataset. An apparent limitation of the research is the need for an ablation study and incomplete hyper-parameters tunning for the pretext task. Besides, due to the hardware limit, we can only roughly afford to train the pretext model, which means failing to ensure its capability to transfer to more complicated tasks. However, the performance will be improved if data engineering is considered and carefully designed in the model for pretext and downstream tasks regarding the same concept presented in this study.

REFERENCES

[1] C. Won, S. Moon and Y. Song, "Open Korean Corpora: A Practical Report," in

Proceedings of Second Workshop for NLP Open Source Software, 2020.

[2] C. Necati Cihan, H. Simon, K. Oscar, N. Hermann and B. Richard, "Neural Sign Language Translation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[3] L. Dongxu, O. Cristian Rodriguez, Y. Xin and L. Hongdong, "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison," in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020.

[4] S. OZGE MERCANOGLU and K. HACER YALIM, "AUTSL: A Large Scale Multi-Modal Turkish Sign Language Dataset and Baseline Methods," IEEE Access, vol. 8, pp. 181340-181355, 2020.

[5] K. Sang-Ki, K. Chang Jo, H. J. and C. Choongsang, "Neural Sign Language Translation Based on Human Keypoint Estimation," Applied sciences, vol. 9, no. 13, p. 2683, 2019.

[6] R. Razieh, K. Kourosh, E. Sergio and S. Mohammad, "Sign Language Production: A Review," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[7] S. Z. Gurbuz, G. Ali Cafer, M. Evie A, G. Darrin J, C. Chris S, R. Mohammad Mahbubur, K. Emre, A. Ridvan, M. Trevor and M. Robiulhossain, "American sign language recognition using rf sensing," IEEE Sensors Journal, vol. 21, no. 3, pp. 3763-3775, 2020.

[8] R. Razieh, K. Kourosh and E. Sergio, "Sign Language Recognition: A Deep Survey," Expert Systems With Applications, vol. 164, p. 113794, 2021.

[9] C. C. Necati, K. Oscar, H. Simon and B. Richard, "Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation," in In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020.

[10] Z. Hao, Z. Wengang, Z. Yun and L. Houqiang, "Spatial-Temporal Multi-Cue Network for Continuous Sign Language Recognition," in Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

[11] K. Yin and R. Jesse, "Better Sign Language Translation with STMC-Transformer," in Proceedings of the 28th International Conference on Computational Linguistics, 2020.

[12] B. Saunders, C. Necati Cihan and B. Richard, "Progressive transformers for end-to-end sign language production," in European Conference on Computer Visio, 2020.

[13] L. Ventura, D. Amanda and G.-i.-N. Xavier, "Can everybody sign now? Exploring sign language video generation from 2D poses," in arXiv preprint arXiv:2012.10941, 2020.

[14] S. Stoll, C. Necati Cihan, H. Simon and B. Richard, "Text2Sign: towards sign language production using neural machine translation and generative adversarial networks," International Journal of Computer Vision, vol. 128, no. 4, pp. 891-908, 2020.

[15] J. Zelinka and K. Jakub, "Neural sign language synthesis: Words are our glosses," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020.

[16] C. Chan, G. Shiry, Z. Tinghui and E. Alexei A, "Everybody dance now," in Proceedings of the IEEE/CVF international conference on computer vision, 2019.

[17] Z. Cao, G. Martinez, T. Simon, S. Wei and Y. and Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[18] E. Park and S. Cho, "KoNLPy: Korean natural language processing in Python," in Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, 2014.

[19] T. Kudo and R. John, "SentencePiece: A simple and language independent subword

tokenizer and detokenizer for Neural Text Processing," in arXiv preprint arXiv:1808.06226, 2018.

[20] A. Vaswani, S. Noam, P. Niki, U. Jakob, J. Llion, N. G. Aidan, K. Łukasz and P. Illia, "Attention is all you need," in Advances in neural information processing systems, 2017.

[21] J. Devlin, C. Ming-Wei, L. Kenton and T. Kristina, "Bert: Pre-training of deep bidirectional transformers for language understanding," in arXiv preprint arXiv:1810.04805, 2018.

[22] Y. Liu, O. Myle, G. Naman, D. Jingfei, J. Mandar, C. Danqi, L. Omer, L. Mike, Z. Luke and S. Veselin, "Roberta: A robustly optimized bert pretraining approach," in arXiv preprint arXiv:1907.11692, 2019.

[23] X. Song, S. Alex, S. Yang, D. Dave and Z. Denny, "Fast wordpiece tokenization," in arXiv preprint arXiv:2012.15524, 2020.

[24] L. Sangah, J. Hansol, B. Yunmee, P. Suzi and S. Hyopil, "KR-BERT: A Small-Scale Korean-Specific Language Model," in arXiv:2008.03979, 2020.

[25] L. Hyunjae, Y. Jaewoong, H. Bonggyu, J. Seongho, M. Seungjai and G. Youngjune, "KoreALBERT: Pretraining a Lite BERT Model for Korean Language Understanding".

[26] S. Edunov, O. Myle, A. Michael and G. David, "Understanding back-translation at scale," in arXiv preprint arXiv:1808.09381, 2018.

[27] D. T. Vu, Y. Gwanghyun, L. Chilwoo and K. Jinyoung, "Text Data Augmentation for the Korean Language," Applied Sciences, vol. 12, no. 7, p. 3425, 2022.

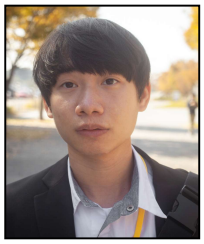
[28] Q. Xie, Z. Dai, E. Hovy, M.-T. Luong and . Q. V. Le, "Unsupervised data augmentation for consistency training," Advances in Neural Information Processing Systems, no. 33, pp. 6256-6268, 2020.

[29] M. Johnson, S. Q. V. L. Mike, K. Maxim, W. Yonghui, C. Zhifeng and T. Nikhil, "Google's multilingual neural machine translation system: Enabling zero-shot translation," in Transactions of the Association for Computational Linguistics, 2017.

[30] N. Reimers and G. Iryna, "Making monolingual sentence embeddings multilingual using knowledge distillation," in arXiv preprint arXiv:2004.09813, 2020.

[31] B. Ban, "A Survey on Awesome Korean NLP Datasets," in arXiv preprint arXiv:2112.01624, 2021.

Authors



Thanh-Vu Dang

He is a student of Ph.D. degree in Department of ICT Convergence System Engineering at Chonnam National University. He received his B.S. degree in Mathematics and Computer Science at Vietnam National University-University of Science, Vietnam in 2018. His research interests are Natural Language Processing, Machine Learning and Deep Learning.

Gwang-Hyun Yu

He is a student of Ph.D. degree in Department of ICT Convergence System Engineering at Chonnam National University. He received his M.S. degree in Electronics Engineering from Chonnam National University, Korea in 2018. His research interests are Digital Signal Processing, Image Processing, Speech Signal Processing, ML, DL.

Ji-yong Kim

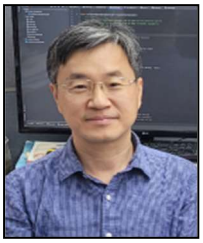
He is a research engineer of Flight Control System at LIG Nex1. He received his B.S. and M.S. degree in Mechatronics Engineering from Chungnam National University, Korea in 2016 and 2019, respectively. He worked in Gwangju Institute of Science and Technology, Korea from 2019 to 2022. His research interests are Exploration robot, Friction Coefficient Estimation, Robot Control, and Korean Sign Language Translation.

Young-Hwan Park

He is a Associate Research Engineer in Satrec Initiative Company, Korea. He received his B.S. and M.S. in Mechatronics Engineering from Chungnam National University, Korea in 2018 and 2021. He worked in Gwangju Institute of Science and Technology, Korea from 2021 to 2022. His research interests are Rough Terrain Slip Control, Robot Engineering, and Sign Language Animation.

Chil-Woo Lee

He is a professor in Department of School of Electronic & Computer Engineering at Chonnam National University, Korea. He received the M.S. degree in Chung Ang University, Korea in 1986 and the Ph.D degree in Tokyo University, Japan in 1992, respectively. His research interest includes Computer Vision, Computer Graphics, Digital Contents and Image Processing.



Jin-Young Kim

He is a professor in Department of ICT Convergence System Engineering at Chonnam National University, Korea. He received his B.S., M.S. and Ph.D. degree in Electronics Engineering from Seoul National University, Korea in 1986, 1988 and 1994, respectively. His research interests are Digital Signal Processing, Image Processing, Speech Signal Processing, ML, DL.