Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1519–1528
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Sign Language Production With Avatar Layering: A Critical Use Case over Rare Words

Jung-Ho Kim[1] **Eui Jun Hwang**[1] **Sukmin Cho**[1] **Du Hui Lee**[2] **Jong C. Park**[1]
[1]KAIST
Daehak-ro 291, Yuseong-gu, Daejeon, Republic of Korea
{jhkim, ejhwang, nelllpic, park}@nlp.kaist.ac.kr
[2]EQ4ALL
Nonhyeon-ro 76-gil 11, Gangnam-gu, Seoul, Republic of Korea
scottlee@eq4all.co.kr

Abstract

Sign language production (SLP) is the process of generating sign language videos from spoken language expressions. Since sign languages are highly under-resourced, existing vision-based SLP approaches suffer from out-of-vocabulary (OOV) and test-time generalization problems and thus generate low-quality translations. To address these problems, we introduce an avatar-based SLP system composed of a sign language translation (SLT) model and an avatar animation generation module. Our Transformer-based SLT model utilizes two additional strategies to resolve these problems: named entity transformation to reduce OOV tokens and context vector generation using a pretrained language model (e.g., BERT) to reliably train the decoder. Our system is validated on a new Korean-Korean Sign Language (KSL) dataset of weather forecasts and emergency announcements. Our SLT model achieves an 8.77 higher BLEU-4 score and a 4.57 higher ROUGE-L score over those of our baseline model. In a user evaluation, 93.48% of named entities were successfully identified by participants, demonstrating marked improvement on OOV issues.

Keywords: sign language, sign language production, signing avatar, Korean, Korean Sign Language

1. Introduction

Sign language (SL) is the primary communication method for the Deaf[1] community. Unlike spoken language, sign language conveys meaning via the movements of hands, face, and body. Moreover, sign language has a different grammar and lexicon from the local spoken language (Sandler and Lillo-Martin, 2006). Therefore, Deaf people usually need a sign language interpreter to interact in hearing society. However, the total need for interpreting services exceeds service availability, especially for translation contexts considered low-priority. Furthermore, access to interpreting services is crucial during disaster and emergency situations where interpreters cannot be requested in advance.

As an alternative, signing avatars have been widely researched for delivering information in sign language (Huenerfauth, 2008; Ebling and Huenerfauth, 2015; Kacorri and Huenerfauth, 2016). Avatar-based animation offers flexibility when creating new expressions but are not well accepted by the Deaf community due to the lack of fluency and naturalness in generated animations. Recently, end-to-end vision-based sign language production (SLP) has received a lot of attention and has shown significant progress (Stoll et al., 2020; Saunders et al., 2021; Hwang et al., 2021). However, it is still challenging to generalize these models to new contexts and accurately reflect nuanced sign variations in sign language expressions. Additionally, pure vision-based SLP is not effective at modeling untrained words, known as the out-of-vocabulary (OOV) problem. In contrast, avatar-based SLP can handle OOV words by programmatically spelling out untrained words, a process that native signers frequently use, called fingerspelling. As OOV words inevitably occur in the real-world setting of SLP, it is necessary either to represent those words properly or to reduce their occurrences. In addition, the lack of parallel data can lead to poorly trained SLP models, producing low-quality sign language video translations.

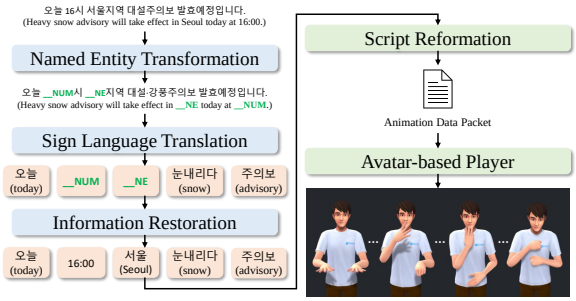

Figure 1: An SLP example via the proposed system.

To resolve these issues in existing SLP approaches, we propose an avatar-based SLP system by combining a sign language translation (SLT) model and an avatar-based animation player. Figure 1 shows an SLP example via the proposed system. Our SLP system first

[1]The uppercase "Deaf" refers to those deaf people who share a sign language and a culture, and the lowercase "deaf" refers to the condition of not hearing (Padden and Humphries, 1989).

translates Korean text into a Korean Sign Language (KSL) gloss sequence. To reduce OOV and generalization problems during this translation, we propose two approaches: (1) transforming named entities and numerical expressions to special tokens in both Korean and KSL sequences, and (2) using a pretrained language model (PLM) as an encoder. After translating the Korean text, the system reformats the KSL gloss sequence into an animation data packet, and then the avatar-based player generates a sign language animation. To validate the system and assess possible sign language interpreting services for critical and disaster situations, we use a new Korean-KSL parallel dataset of weather forecasts and government emergency announcements.

The main contributions of our work to the field of SLP are summarized as follows.

- We introduce a new Korean-KSL parallel corpus of weather forecasts and emergency announcements for SLT and SLP models. The full dataset will be released in late 2022.
- We propose two methods to better translate untrained words and quantitatively demonstrate the effectiveness of the proposed methods.
- We present an avatar-based animation system and validate its performance via user evaluation.

The rest of this paper is structured as follows. Section 2 discusses related work on SLT and SLP. Section 3 introduces the Korean-KSL weather forecasts and emergency announcements dataset. Sections 4 and 5 describe the two SLT methods and the avatar-based SLP method, respectively. Section 6 describes the experimental setup and presents a detailed analysis of the results. Section 7 concludes this study and suggests future work.

2. Related Work

2.1. Sign Language Translation

Early studies on SLT are mostly based on statistical approaches. Bungeroth and Ney (2004) formulate an SLT task as a text-to-text translation problem by representing sign language expressions as a sequence of glosses. Morrissey et al. (2010) propose an SLT model that translates from English to Irish Sign Language (ISL), German to ISL, English to German Sign Language (DGS), and German to DGS. Stein et al. (2012) propose an SLT model from German to DGS using the RWTH-PHOENIX-Weather-2012 dataset, which covers the weather forecast domain.

With the emergence of sequence-to-sequence learning, neural machine translation (NMT) has become an active area of research. SLT researchers have henceforth adopted many NMT techniques and applied them to SLT. However, since there was no large-scale training dataset for SLT, it was impossible to achieve satisfactory performance using an encoder-decoder model until Camgöz et al. (2018) released such a dataset for SLT. They also proposed an RNN-based encoder-decoder model to translate sign videos to sign language gloss sequences. Moryossef et al. (2021) proposed a data augmentation method based on lexical overlap between spoken and sign languages to address the performance degradation of NMT models due to the scarcity of SL data. Yin and Read (2020) utilized the Transformer model architecture with various encoding and decoding schemes to translate from a sequence of sign language glosses to a spoken language sequence and analyzed the impact of the proposed schemes on translation performance.

Despite these efforts, the methods above still have a clear limitation. NMT models often generate poor translations when their training dataset contains many rare (or low-frequency) words. According to (Li et al., 2021), the problem could be exacerbated when there is not enough data to learn meaningful correlations or when the target sentences are morphologically rich and complex. Tu et al. (2012) point out the difficulties in translating numerical expressions. Recently, Wang et al. (2021) also argue that numerical text mistranslation is a general problem. Through behavioral testing, they demonstrated that even major commercial systems and state-of-the-art NMT models fail on many numerical text-related tasks. Therefore, it is necessary to explore translation methods better designed for numerical expressions and other types of rare words (such as person's names, locations, organizations, etc.) to generate more accurate translations.

2.2. Sign Language Production

Signing avatars were initially investigated for automatic generation of new sign language content in a web-based environment. At their initial stage, researchers focused on specifying hand and arm movements (Lebourque and Gibet, 1999; Gibet et al., 2001). To fully convey the meaning of sign language expressions, studies have been conducted to reflect non-manual features such as facial movement and space utilization (Huenerfauth, 2008; Kipp et al., 2011; Schmidt et al., 2013; Adamo-Villani and Wilbur, 2015; Kacorri and Huenerfauth, 2016). Sign language notation systems such as HamNoSys (Prillwitz, 1989), SignWriting (Sutton, 2014), and SigML (Elliott et al., 2004) have been used to define functionalities of avatar animation scripts. Though a drawback of avatar-based SLP has been the lack of realism and expressiveness of avatars, recent datasets that include motion capture data have been published and exploited to produce more realistic sign language movements using avatars (Brock and Nakadai, 2018; Naert et al., 2020).

Recently, vision-based, or data-driven, SLP has become a growing research field. Stoll et al. (2018) first propose an NMT-based SLP model consisting of an SLT model and a generative model to produce realistic sign language videos. To augment their own SLP

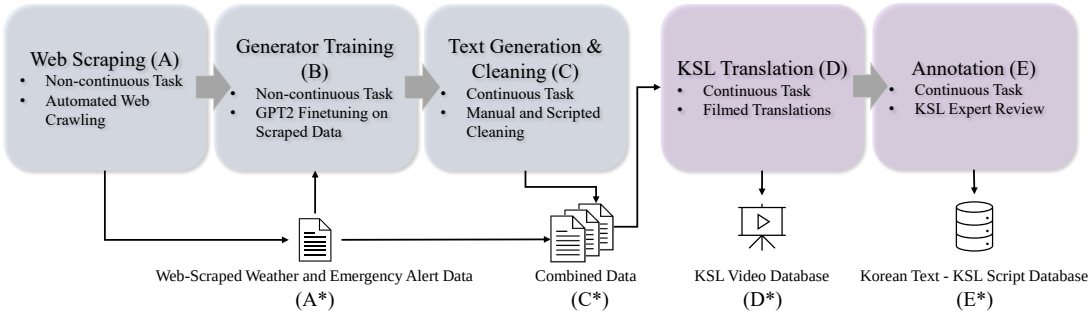

Figure 2: An overview of the data collection method. Data collection can be broken into five steps: Web scraping for Korean text (A), Korean text generator training (B), Korean text generation (C), KSL translation (D), and KSL Annotation (E). Step A is a one-time task that produces a set of Korean weather and emergency alert messages (A*). This data is used to train the generators in step B, also a one-time task. Additional data is then generated and cleaned in step C and combined with A* to make C*. This step is continually performed for each weather and alert subcategory. In step D, each data instance in C* is translated, creating the KSL video database (D*). Finally, each video in D* is annotated by native or expert signers and collected into a parallel corpus database (E*).

model, Zelinka and Kanis (2020) presented a gradient-descent-based method for skeletal model estimation refinement that can smooth bone positions, interpolate missing bones, and expand 2D bone markers to 3D. Saunders et al. (2020) proposed a transformer-based SLP model that can generate variable-length sign pose data using a counter encoding scheme. Saunders et al. (2021) leveraged gloss labels to learn motion primitives alongside a mixture of experts (MoE) model to generate translations with high user ratings. Hwang et al. (2021) proposed a non-autoregressive SLP model to prevent joint position predictions from regressing to the mean.

3. Corpus Construction

In this section, we introduce a new large-scale Korean-KSL parallel corpus of weather forecasts and government emergency announcements. We explain how the corpus was constructed and report its statistics. Note that the corpus used in this research is a subset of a corpus from a separate ongoing data generation project, and the full corpus will be publicly released in late 2022. For the rest of this paper, we will refer to the organizations collaborating on this data project collectively as "data collectors" and use the past tense to discuss the data collection project since the collection of our subset has already been completed.

3.1. Corpus Overview

The goal of the data generation project was to create a Korean-KSL parallel corpus for training a machine translation model, specifically over Korean weather report and government emergency announcement domains. Each data instance contains one Korean passage (usually segmented at the sentence level) and one KSL translation (represented by a video and an associated annotation). Emergency announcements fall into multiple categories, such as typhoon, flooding, heavy rain, heavy snow, financial system failure, and terrorism.

Data was collected using a five-step process: (1) original text collection, (2) text generator training, (3) ML-based text augmentation, (4) KSL translation, and (5) KSL annotation. Native signers were deeply involved in the translation and annotation process. Figure 2 shows an overview of the data collection procedure.

Unlike most parallel corpora, this dataset is pre-augmented – data collectors generated additional Korean text using learned generative models during the initial Korean text collection phase. This has the benefit of allowing fine control over class imbalance, avoiding the manual generation of new weather reports and emergency announcements, and allowing for the creation of high-quality KSL augmentations. However, generated data showed qualitatively low variance, which may lead to artificially inflated validation performance.

3.2. Korean Text Collection

Data collectors first scraped the web for Korean weather reports and emergency announcements. Edwards et al. (2021) demonstrated the effectiveness of using generative models (in their case, GPT-2 (Radford et al., 2019)) to augment data for classification tasks. Similarly, data collectors used the scraped data to train a series of generate language models which were leveraged to generate a large pool of augmented Korean text data. This pool contained significant noise and required thorough cleaning. Noisy instances were edited or removed, and redundant instances and instances deemed to be outside of the target distribution were also removed. Data size across categories was also rebalanced (i.e., oversampled) through this step. Special symbols, such as ♣ and ►, are used frequently in Korean government emergency announcements. Data collectors chose to not remove these special symbols, allowing potential dataset users full control over pre-processing. Finally, both scraped and augmented sentences were combined to generate the Korean source dataset.

Figure 3: An excerpt from the annotation tool web page. Each horizontal annotation tier represents a different annotation type. The first tier is the source Korean text and is automatically filled in. Respectively, the second and sixth annotation tiers are for dominant hand and non-dominant hand gloss and FS annotations. All tiers from the eighth and below are for NMS annotations. The subject's face in this screenshot is masked for privacy.

3.3. Korean Sign Language Collection

Before recording translations, each Korean passage (see §3.2) was pre-translated into KSL by a native signer. This translation was usually prerecorded, but some translators chose to simply memorize or transcribe the intended translation. Sometimes translators (or collaborating KSL experts) judged that a Korean text should have multiple distinct KSL translations. In these cases, a single Korean text could have up to three separate translations prepared. Each unique translation was treated as a new data instance and logged separately. As a result, the dataset includes Korean text repetition (and possible KSL repetition), but each instance is a unique Korean-KSL pair.

Translations were officially filmed either in one of several studios against a neutral background or crowdsourced out to native signers in the community. Data collectors could not control filming conditions for crowdsourced videos, but significantly more videos could be generated using crowdsourcing. All participating translators were Deaf, and both hearing KSL experts and several Deaf signers annotated the translations. A total of 139 people participated in translating, filming, and annotating KSL data instances.

The KSL videos were then annotated using a custom web app (EQ4ALL, 2022a) for multi-modal video annotation (see Figure 3). Signs, non-manual signals (NMS), and fingerspellings (FS) were identified and segmented. Depending on the annotation entry, additional information (such as position) was also annotated.

3.4. Parallel Corpus Format

Each data instance is represented by a KSL MP4 video file and a JSON file containing Korean text and KSL annotations. Each JSON file has the following four sections:

- Metadata: information related to the creation and annotation of the data instance, including a flag identifying instances using augmented Korean text data.
- Korean Text: the source Korean sentence.
- Manual Sign List: three separate gloss lists – one for signs or FS performed by the dominant hand, one for signs or FS performed by the non-dominant hand, and one for signs or FS that require both hands and are performed together. Sign and FS start and end times are included for each entry.
- Non-Manual Sign List: separate lists for each type of annotated NMS. NMS start and end times are included for each entry.

3.5. Statistics

After removing instances with repeated Korean source texts (see explanation on Korean sentence repetition in §3.3), the dataset consists of 22,875 unique Korean-KSL data instances. Korean text was processed using a pretrained tokenizer with a vocabulary size of 4,117, as explained in §6. The processed Korean text data has 716,871 tokens. The KSL annotation data has 419,688 total segmentations, with a total of 7,949 unique glosses.

All data instances were classified into one of two main categories: weather or emergency announcements. However, emergency announcements were further grouped into one of thirty-seven subcategories (though some announcements are qualitatively cross-category). Subcategories are related to natural disasters, industrial accidents, dangerous activities, health, and service or infrastructure outages.

4. Sign Language Translation

In this section, we propose two different SLT techniques for KSL: named entity transformation and using a PLM as an encoder. Note that the two SLT methods can be used together. Figure 4 shows an overview of our SLT system.

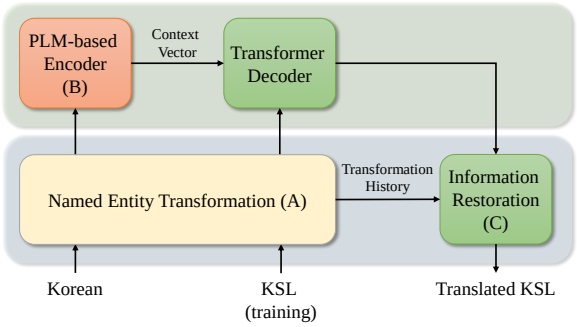

Figure 4: An overview of our SLT system. First, the named entity transformation module transforms input sentences (see (A)). Next, a PLM (e.g., BERT) encodes the transformed source sentences to context vectors (see (B)) and the decoder generates translated KSL sequences containing named entity tokens. Finally, the information restoration module converts named entity tokens to their original expressions using a transformation history (see (C)).

4.1. Problem Definition

Let X denote a source word sequence. The goal of SLT is to produce a corresponding target gloss sequence $Y = (y_1, ..., y_M)$ for each $x \in X$ with tokens $(x_1, ..., x_N)$, which is a maximization task of a probability distribution $P(Y|X)$.

4.2. Baseline: Transformer

We employ the Transformer (Vaswani et al., 2017), a de-facto standard model in NMT, as a baseline. It is based solely on attention mechanisms to generate representations of entire sequences with global dependencies. Attention can be computed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \quad (1)$$

where Q, K and V are query, key and value, respectively. Multi-head attention (MHA) layers use the equation (1) to model different weighted combinations of each input sequence. This allows the model to perform a powerful analysis of its inputs. MHA can be represented as follows:

$$MHA(Q, K, V) = Concat(head_i, ..., head_n)W_O, \quad (2)$$

$$head_i = Attention(QW_Q^i, KW_K^i, VW_V^i), \quad (3)$$

where W_O,W_Q, and W_V are weights related to each input. Moreover, the model learns the mapping between the source and target sequences using encoder-decoder attention.

4.3. Named Entity Transformation

Spoken and sign languages used in the same country often share a lexicon and letters. In this section, we propose KSL-specific heuristic transformation approaches that exploit this lexical overlap to enhance the KSL translation. Specifically, we suggest a better way for SLT models to learn generic named entities such as names of persons, locations, organizations, and times and quantities (Nadeau and Sekine, 2007).

Original	
Korean	서울지역 대설주의보 발효예정입니다. (Heavy snow advisory will take effect in Seoul.)
KSL	서울 눈내리다 주의보 (Seoul snow advisory)
Transformed	
Korean	_NE지역 대설주의보 발효예정입니다. (Heavy snow advisory will take effect in _NE.)
KSL	_NE 눈내리다 주의보 (_NE snow advisory)

Table 1: An example of named entity transformation for location.

4.3.1. Persons, Locations, and Organizations

We transform all words that need to be fingerspelled into a special token "_NE" in both Korean and KSL sequences. These transformed tokens will be restored after the translation process (see §4.5). By introducing this transformation, an SLT model can support fingerspelled words. Table 1 shows a transformation example.

4.3.2. Times and Quantities

Temporal and numerical expressions must be learned effectively with extremely little data because SLT models treat unseen number combinations as unseen words. In this study, we propose a unifying method to map temporal and numerical expressions into regular expression patterns. We do so by manually analyzing source Korean sequences and target KSL sequences in our parallel corpus to identify representative expressions. By defining the representative expressions as regular expressions, our SLT system can automatically transform any Korean and KSL sequences containing supported temporal and numerical expressions. Table 2 summarizes the supported representative expressions for each type with examples.

With these transformations, SLT models have more opportunities to be trained on the same patterns and thus they become more robust in translating temporal and numerical expressions. Since the proposed transformations in this study are designed based on our training corpus, they do not cover all temporal and numerical expressions in Korean and KSL. We plan to add supported regular expressions as we expand the corpus in the future.

4.4. Using a PLM as an Encoder

A PLM such as BERT (Devlin et al., 2019) can help encode Korean sentences much better than an untrained Transformer encoder when the training dataset is small (Miyazaki et al., 2020). Therefore, we employ a PLM to encode Korean sentences. Although there are

Types	Representative Expressions	Examples	
		Original	Transformed
Numbers	_NUM	-3, 0.4, 10000, 삼십 (thirty)	_NUM
Numbers with units	_NUM(unit)?(_NUM(unit)?)+	7시 30분 (7:30) 3월 7일 (March 7)	_NUM 시 _NUM 분 (_NUM hour _NUM min) _NUM 월 _NUM 일 (_NUM month _NUM day)

Table 2: Representative expressions per number type with examples.

no relevant KSL PLMs that can be used in place of the decoder, utilizing the output sequence of the encoder z_{BERT} allows for effective decoder training.

4.5. Information Restoration

When an SLT model takes Korean sentences containing transformed tokens as input, it should produce transformed tokens in the translated KSL sequences. As the final step, the information restoration module determines the relationship between source-side named entity tokens and target-side named entity tokens by using the cross attention of the Transformer decoder and then restores target-side named entity tokens to their original words by using transformation history received from the named entity transformation module (see §4.3).

5. Avatar-based SLP

Our avatar-based SLP method has three main steps: (1) first the SLT model is utilized to create a KSL translation from an input Korean text; (2) this translation is then reformatted into an animation data packet; and (3) finally the data packet is fed to the animation player, which produces a complete sign language animation via avatar layering.

5.1. Animation Data Packet Format

Our animation data packet is time-based and multi-channel. For a single data packet (representing one KSL translation), each channel contains a list of animations and their associated animation data, combined into animation "blocks". Each block should detail the animation type, animation ID, animation start and end times, and additional fields depending on the animation type. In general, each block corresponds to exactly one sign, gesture, or NMS, but blocks can also correspond to multiple short signs blended together or sign sub-units that need to be combined with other animations. Our script has the following animation channels:

- SequenceHandBoth: animations using both hands
- SequenceHandRight: right hand animations
- SequenceHandLeft: left hand animations
- SequenceFace: animations for facial NMS (smile, raised eyebrows, etc.)
- SequenceMouth: jaw-based NMS animations (mouth opening, jaw shaking back and forth, etc)

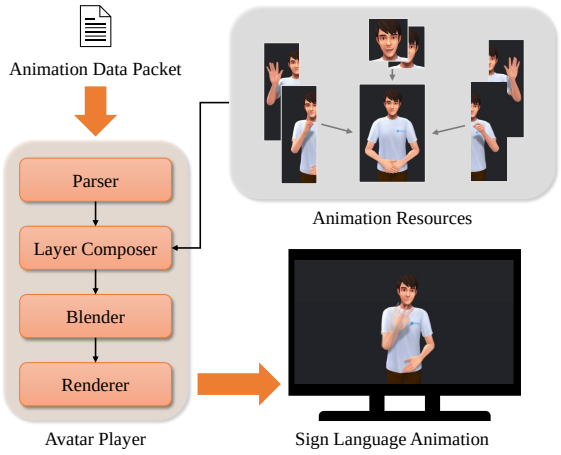

Figure 5: A high-level overview of our avatar player pipeline. The player parses the input data packet, composes animation layers based on each input channel, calculates animation timing and inter-animation blending data, and renders the avatar video.

- SequenceBody: animations for general body NMS (head movement, shoulders, etc.)

We found that utilizing a joint right & left hand channel (SequenceHandBoth) for signs and gestures involving both hands made it easier to generate natural looking signs.

Note that, though there are channels for face- and mouth-based NMS (SequenceFace and SequenceMouth, respectively), only body-based NMS (specifically, head movements) are implemented in our current avatar player, and further development is needed to effectively integrate all NMS channels. Furthermore, note that our SLT model was not trained on NMS labels and did not use the SequenceHandRight or SequenceHandLeft channels for the experiments in §6. We recognize this as a limitation of our study and will improve our SLT model in the future.

5.2. Player Pipeline

Our avatar player (EQ4ALL, 2022b) draws from a pool of avatar animation resource files during video generation. The avatar player pipeline can be summarized in the following steps. On receiving a data packet (see §5.1), the player (1) parses sign, NMS, and FS blocks from each data channel; (2) composes animation

layers according to the parsed channel-wise data; (3) calculates animation timing and inter-animation blending data; and (4) renders an avatar video (see Figure 5). By decomposing animations into separate channels and layering them, we are able to increase the number of sign, sub-sign, fingerspelling, and NMS combinations that we can produce. With this increased gesture range, the challenge becomes layering and blending motions in a natural way, which is an area we are actively researching.

6. Experiments

In this section, we describe the experimental setup and report experimental results for our SLP system.

6.1. Experimental Setup

6.1.1. Dataset

We split the parallel corpus into three datasets: training, validation, and test datasets, with a ratio of 90:5:5.
Table 3 shows the key statistics of training, validation, and test datasets.

6.1.2. Metrics

To assess the quality of SLT in our SLP system, we employ two widely used metrics, BLEU (Papineni et al., 2002) and ROUGE-L F_1 (Lin, 2004). For BLEU, we calculate BLEU-1 through BLEU-4 to better evaluate word-level and phrase-level translations.

6.1.3. Implementation Details

For our baseline and named entity transformation models, we use a Transformer model with four encoder layers and four decoder layers. The embedding size is set to 768 and the number of attention heads is set to 8. For the SLT model using a PLM as an encoder, we use the KLUE-BERT-base (Park et al., 2021) as an encoder and an untrained Transformer decoder with six layers and eight attention heads. We train all models for 20 epochs with early stopping and use AdamW optimization (Loshchilov and Hutter, 2019) starting with a learning rate of 10^{-4}. Note that we also train the PLM encoder together as Korean sentences may contain named entity tokens on which the PLM was not trained.

6.2. Experimental Results

In this subsection, we describe experimental results that demonstrate the superiority of our proposed models over the baseline Transformer. We report quantitative and qualitative results in the following subsections. In all experimental results, we respectively denote SLT models using named entity transformation (see §4.3) and using a PLM as an encoder (see §4.4) as "A" and "B." We further denote the combination of both methods as "A+B."

6.2.1. Impact of the Proposed Methods

Table 4 summarizes the quantitative performance of all models. Models A and B outperform the baseline Transformer by 6.20 and 2.05 BLEU-4 scores, respectively. The combined SLT model (A+B) achieves 41.46 BLEU-4 score and also outperforms the baseline by 8.77 BLEU-4 score. ROUGE-L scores showed similar results. Overall, the combined SLT model shows better performance than the two individual SLT models, and all models improve on the baseline Transformer model. We could expect some performance degradation when embedding a Korean sentence containing transformed tokens (_NE and _NUM) with the PLM since they are untrained tokens. However, the net benefit of using a PLM seems to outweigh any possible degradation.

	Korean			KSL		
	Train	**Val**	**Test**	**Train**	**Val**	**Test**
# Sequence	21,415	730	730	21,415	730	730
# Words*	670,687	23,100	23,084	392,887	13,351	13,450
Vocab Size	4,016	1,680	1,661	7,727	1,703	1,680
# OOV words	-	37	64	-	105	118

Table 3: Statistics of training, validation, and test datasets. *The number of words is calculated by tokenizing sentences using the same PLM tokenizer for our SLT models.

6.2.2. User Evaluation

We also conducted user evaluation to assess the quality of our avatar-based SLP system. We recruited four participants who use KSL as their primary language. We then randomly sampled 30 sign language animation videos from the production results on our test set.
We asked the following three questions for each sign language animation.

Q1. Naturalness (1-5): How natural is the expression in the sign language animation?

Q2. Accuracy (1-5): How accurately does the sign language animation match the original video?

Q3. Named entity identification: Is there a location name or numerical expression in the sign language animation? If there is, please write it in Korean.

Note that all three questions were asked using KSL supplemented by written prompts. Furthermore, Q2 is target-based direct assessment and is asked after showing the participants the original translation video as a reference.
Naturalness and accuracy are scored on a 5-point Likert scale ranging from 1 (most negative) to 5 (most positive). Figure 6 shows naturalness and accuracy scores of our animation results per category. Overall, our SLP system achieves about 2.27 scores for both criteria. According to the comments from the participants, the main reason for these scores is the unnecessary repetition of signs in the translated KSL sequence. Our SLT model often generates repeated KSL glosses. This is a known problem with beam search (Fan et al., 2018; Holtzman et al., 2019), the decoding algorithm used for translation generation in all experiments. Furthermore,

Model	Validation					Test				
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L
Baseline	64.42	51.36	41.88	34.78	62.18	63.31	49.68	39.95	32.69	61.46
A	68.25	55.59	46.08	38.89	65.33	68.54	55.76	46.12	38.89	65.12
B	66.09	53.51	44.06	36.89	64.08	64.67	51.65	42.08	34.74	63.09
A+B	**69.05**	**56.58**	**47.14**	**40.02**	**66.23**	**69.35**	**57.44**	**48.37**	**41.46**	**66.03**

Table 4: Translation scores for all models. The figure in bold indicates the top score by the respective metric.

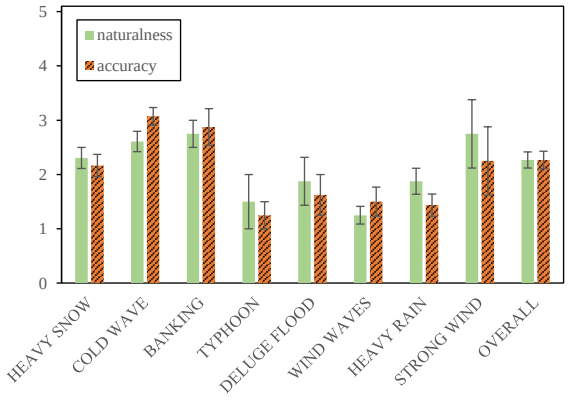

Figure 6: Naturalness and accuracy scores per category

	Naturalness	Accuracy
BLEU-4	0.3544 (p-value < 0.1)	0.5491 (p-value < 0.01)

Table 5: Pearson correlation coefficients between sentence-level BLEU-4 score and human evaluation scores.

	Location	Time	Total
Identification Rate (%)	95.45	91.67	93.48

Table 6: Average identification rates by participants for named entities

since KSL is inherently a spatial language, finding an appropriate decoding scheme is still an open problem. We plan to address these issues in future research.

Inter-Rater Reliability (IRR) To determine the agreement among participants, we calculated Fleiss' kappa coefficient (κ) (Fleiss, 1971). Both IRRs for naturalness and accuracy show substantial agreements ($\kappa_{\text{naturalness}} = 0.625$ and $\kappa_{\text{accuracy}} = 0.657$).

Correlation with Translation Scores We also measured how much user evaluation scores are correlated with the sentence-level BLEU-4 score by calculating Pearson's correlation coefficient (PCC, ρ). Table 5 shows the PCCs ($\rho_{\text{naturalness}}$ and ρ_{accuracy}) between sentence-level BLEU-4 score and human evaluation scores, respectively. According to the PCC interpretation in (Dancey and Reidy, 2007), $\rho_{\text{naturalness}}$ shows a weak positive correlation, while ρ_{accuracy} indicates a moderate positive correlation. Despite these correlations, there is a gap between absolute quantitative scores (41.46 BLEU-4 and 66.03 ROUGE-L, both considered high scores) and user evaluation scores (2.27 naturalness and accuracy scores). We draw the following two conclusions from this. First, translations with a higher BLEU score are qualitatively better translations and BLEU is effective for ranking translations. Second, there is a clear need for new metrics for evaluating SLP models to be developed. Conventional metrics cannot evaluate spatial and non-manual components of a sign language translation.

Named Entity Identification We measured how well participants identify named entity expressions in sign language animations. Table 6 shows the identification rates per subcategory of named entities. All participants recognized named entity expressions with high precision. Viewed against our SLT model's strong results in the previous sections, this result demonstrates that the proposed SLP system can effectively translate Korean text into KSL animations without learning individual named entity expressions.

7. Conclusion and Future Work

We introduced a Korean-KSL parallel corpus of weather forecasts and emergency announcements and presented an avatar-based SLP system that combines an SLT model with a signing avatar. To alleviate OOV issues in SLP, we improved our SLT models by transforming named entities into special tokens and using a PLM as an encoder. Through the experiments, we obtained high translation scores (41.46 BLEU-4 and 66.03 ROUGE-L), though there is still some gap with user evaluation scores. Nevertheless, our SLP system achieved a high identification rate (93.48%) for named entities in sign language animations, demonstrating the improved processing of OOV words.

For future work, we will generalize our proposed model to other languages. We also plan to look into a sequence-to-graph SLT model exploiting the linguistic nature of sign language to better utilize our signing avatar mechanism and work on a new metric for evaluating SLT models (see §6.2.2).

8. Acknowledgments

This work was supported by the Technology Innovation Program (20014406, Development of interactive sign language interpretation service based on artificial intelligence for the hearing impaired) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea).

9. Bibliographical References

Adamo-Villani, N. and Wilbur, R. B. (2015). Asl-pro: American sign language animation with prosodic elements. In *International Conference on Universal Access in Human-Computer Interaction*, pages 307–318. Springer.

Brock, H. and Nakadai, K. (2018). Deep jslc: A multimodal corpus collection for data-driven generation of japanese sign language expressions. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bungeroth, J. and Ney, H. (2004). Statistical sign language translation. In *Workshop on representation and processing of sign languages, LREC*, volume 4, pages 105–108. Citeseer.

Camgöz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural Sign Language Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.

Dancey, C. P. and Reidy, J. (2007). *Statistics without maths for psychology*. Pearson education.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ebling, S. and Huenerfauth, M. (2015). Bridging the gap between sign language machine translation and sign language animation using sequence classification. In *Proceedings of SLPAT 2015: 6th workshop on speech and language processing for assistive technologies*, pages 2–9.

Edwards, A., Ushio, A., Camacho-Collados, J., de Ribaupierre, H., and Preece, A. (2021). Guiding generative language models for data augmentation in few-shot text classification.

Elliott, R., Glauert, J. R., Jennings, V., and Kennaway, J. (2004). An overview of the sigml notation and sigmlsigning software system. In *Fourth International Conference on Language Resources and Evaluation, LREC*, pages 98–104.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gibet, S., Lebourque, T., and Marteau, P.-F. (2001). High-level specification and animation of communicative gestures. *Journal of Visual Languages & Computing*, 12(6):657–687.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Huenerfauth, M. (2008). Generating american sign language animation: overcoming misconceptions and technical challenges. *Universal Access in the Information Society*, 6(4):419–434.

Hwang, E. J., Kim, J.-H., and Park, J. C. (2021). Non-Autoregressive Sign Language Production with Gaussian Space. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Kacorri, H. and Huenerfauth, M. (2016). Selecting exemplar recordings of american sign language nonmanual expressions for animation synthesis based on manual sign timing. In *Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies, INTERSPEECH*.

Kipp, M., Heloir, A., and Nguyen, Q. (2011). Sign Language Avatars: Animation and Comprehensibility. In *International Workshop on Intelligent Virtual Agents*, pages 113–126. Springer.

Lebourque, T. and Gibet, S. (1999). High level specification and control of communication gestures: the gessyca system. In *Proceedings Computer Animation 1999*, pages 24–35. IEEE.

Li, J., Shen, Y., Huang, S., Dai, X., and Chen, J. (2021). When is char better than subword: A systematic study of segmentation algorithms for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 543–549.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out, Post-Conference Workshop of ACL 2004*, pages 74–81.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

Miyazaki, T., Morita, Y., and Sano, M. (2020). Machine translation from spoken language to sign language using pre-trained language model as encoder. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 139–144.

Morrissey, S., Somers, H., Smith, R., Gilchrist, S., and Dandapat, S. (2010). Building a sign language corpus for use in machine translation. In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 172–177.

Moryossef, A., Yin, K., Neubig, G., and Goldberg, Y. (2021). Data augmentation for sign language gloss translation. *arXiv preprint arXiv:2105.07476*.

Nadeau, D. and Sekine, S. (2007). A survey of named

entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Naert, L., Larboulette, C., and Gibet, S. (2020). Lsf-animal: a motion capture corpus in french sign language designed for the animation of signing avatars. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6008–6017.

Padden, C. and Humphries, T. (1989). Deaf in america: Voices from a culture. *Ear and Hearing*, 10(2):139.

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

Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., Song, C., Kim, J., Song, Y., Oh, T., et al. (2021). Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Prillwitz, S. (1989). *HamNoSys Version 2.0. Hamburg Notation System for Sign Languages: An Introductory Guide*. Intern. Arb. z. Gebärdensprache u. Kommunik. Signum Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Sandler, W. and Lillo-Martin, D. (2006). *Sign language and linguistic universals*. Cambridge University Press.

Saunders, B., Camgöz, N. C., and Bowden, R. (2020). Progressive Transformers for End-to-End Sign Language Production. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Saunders, B., Camgöz, N. C., and Bowden, R. (2021). Mixed signals: Sign language production via a mixture of motion primitives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1919–1929, October.

Schmidt, C., Koller, O., Ney, H., Hoyoux, T., and Piater, J. (2013). Enhancing gloss-based corpora with facial features using active appearance models. In *International Symposium on Sign Language Translation and Avatar Technology*, volume 2, pages 41–49. Chicago, IL, USA.

Stein, D., Schmidt, C., and Ney, H. (2012). Analysis, preparation, and optimization of statistical sign language machine translation. *Machine Translation*, 26(4):325–357.

Stoll, S., Camgöz, N. C., Hadfield, S., and Bowden, R. (2018). Sign Language Production using Neural Machine Translation and Generative Adversarial Networks. In *Proceedings of the 29th British Machine Vision Conference (BMVC)*. British Machine Vision Association.

Stoll, S., Camgöz, N. C., Hadfield, S., and Bowden, R. (2020). Text2sign: Towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4):891–908.

Sutton, V. (2014). *Lessons in sign writing: Textbook*. SignWriting.

Tu, M., Zhou, Y., and Zong, C. (2012). A universal approach to translating numerical and time expressions. In *Proceedings of the 9th International Workshop on Spoken Language Translation: Papers*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u., and Polosukhin, I. (2017). Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010.

Wang, J., Xu, C., Guzmán, F., El-Kishky, A., Rubinstein, B. I., and Cohn, T. (2021). As easy as 1, 2, 3: Behavioural testing of nmt systems for numerical translation. *arXiv preprint arXiv:2107.08357*.

Yin, K. and Read, J. (2020). Better sign language translation with stmc-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989.

Zelinka, J. and Kanis, J. (2020). Neural Sign Language Synthesis: Words are our Glosses. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3395–3403.

10. Language Resource References

EQ4ALL. (2022a). *EQ4ALL Annotation Tool*.

EQ4ALL. (2022b). *EQ4ALL Avatar Player*.