JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 25, No. 10, pp. 2829-2840, Oct. 2024

Check for updates

농인 중심의 기계번역을 위한 한국수어 스크립트 제작 및 텍스트 데이터 전처리 연구

주 용 민 · 김 소 진*
광주과학기술원 한국문화기술연구소 연구원

Study of the Production of Korean Sign Language Scripts and Preprocessing of Text Data for Machine Translation Targeting Deaf Individuals

Yong-Min Ju · So-Jin Kim*
Researcher, Korea Culture Technology Institute, Gwangju Institute of Science and Technology, Gwangju 61005, Korea

[요 약]

인공지능 자연어처리 기술 발전에 따라 최근 번역시스템들은 기존보다 높은 수준의 번역을 할 수 있다. 그러나 한국수어 같은 특수영역에서는 여전히 자동화된 번역시스템을 찾아보기 어렵다. 본 연구에서는 이에 대해 기계번역 관점에서 대형언어모델을 학습시킬 수 있는 수어 데이터가 부족한 것에 원인이 있다고 가정하였다. 이를 해결하기 위해 분산 되어있는 수어 데이터들을 수집하고, 개별 데이터셋의 공통 규칙을 발견하여 기계학습에 적용할 수 있는 수준의 일관성 있는 훈련 데이터로 전처리하기 위한 다중 코퍼스 텍스트 데이터 전처리 알고리즘을 개발하였다. 또한 코드 변환 과정단계를 통해 한국수어 애니메이션을 생성할 수 있는 데이터 스크립트를 제작하였다. 본 연구를 통해 인력 소모가 심한 텍스트 데이터 전처리 비용을 획기적으로 단축할 수 있었으며, 자체 실험 결과 평균 87.6%의 정확도로 수어 텍스트 데이터를 게임엔진에서 구동할 수 있는 코드로 변환할 수 있었다. 이러한 과정으로 데이터셋이 적은 저자원의 특수 환경의 데이터셋 처리 파이프라인 및 프로세스에 관한 효율적 방안을 제시하였으며, 한국수어 번역을 위한 기계번역 성능 고도화 가능성을 확인하였다.

[Abstract]

With the development of AI-based natural language processing, machine translation systems have surpassed human translation in some areas. However, effective translation systems for specialized fields, such as Korean sign language, remain rare. Due to the scarcity of Korean sign language data for training machine translation models, we first gathered distributed sign language data and identified common patterns within the dataset. Next, we developed a multi-corpus text preprocessing algorithm to standardize the training data for machine learning. Additionally, we created a data script to generate Korean sign language animations, incorporating a code conversion process. This approach reduced labor-intensive preprocessing costs and enabled the conversion of sign language text into executable code with an average accuracy of 87.6%. This study presents a solution for processing Korean sign language datasets for machine translation, even with limited data resources.

색인어 : 데이터 전처리, 한국수어, 머신러닝, 기계번역, 자연어처리
Keyword : Data Preprocessing, Korean Sign Language, Machine Learning, Machine Translation, Natural Language Processing

http://dx.doi.org/10.9728/dcs.2024.25.10.2829

Received 12 August 2024; **Revised** 25 September 2024
Accepted 02 October 2024

***Corresponding Author; So-Jin Kim**
Tel: +82-62-715-4928
E-mail: somolojin@gist.ac.kr

Ⅰ. 서 론

1-1 연구 배경

최근의 한국수어와 관련된 연구들은 2016년 개정된 한국수화언어법에 의거하여 한국수어가 한국어와 함께 대한민국의 공용어로서의 언어적으로 대등한 위치로 지정되었다는 것, 그리고 지금까지의 기술 수준을 뛰어넘는 기계학습 자연어 처리 분야의 발전에 따라 개발된 ChatGPT(Chat Generative Pre-Transformer)와 같은 인공지능 신모델들의 등장 등의 변화로 인하여 다수의 연구자들에게 많은 관심을 불러일으키고 있다.

초거대 인공지능 모델의 학습을 위해 사용되는 시퀀스-투-시퀀스(Sequence-to-Sequence), 트랜스포머(Transfomer), 그리고 버트(Bidirectional encoder representations from transformers, BERT)와 같은 알고리즘에 기반한 고도의 자연어처리 모델들은 최근 어텐션(Attention) 기법들을 복합적으로 구성하여 영상처리 및 음성처리 분야에서 적용되어 다양하게 활용되고 있다[1]. 수어 기계번역 연구에서도 이를 적용하여 음성적 표현을 매개로 하는 청인과 비음성적 표현인 시각을 기반으로 의사소통을 하는 농인간의 커뮤니케이션을 위해 한국수어 기계학습 번역모델을 개발하고 모델의 성능을 고도화하여 수어 번역 성과를 높이기 위한 연구들이 진행 중이다[2].

현재 이러한 주요 연구들은 한국수어를 한국어로 인식하여 농인의 언어를 청인에게 전달하는 단 방향성 연구가 중점적으로 수행되고 있으며, 특히 컴퓨터 비전 기술을 중심으로 영상처리 분야에서 다수의 연구가 수행되고 있다[3]. 이는 한국의 수어 연구뿐만 아니라 국가별 수어 연구 동향에서도 나타나는 전 세계적 흐름으로 국내의 연구 방향과 유사한 형태를 보인다. 수어 영상과 이미지를 인식하고 한국수어 텍스트로 변환하는 과정에서 발생하는 저성능의 수어 인식 문제를 별건으로 취급하더라도 영상 및 이미지를 텍스트 형태의 정답으로 도출하는 청인이 수어를 이해하기 위한 일방향성 연구만으로는 농인과 청인 상호의 의사소통을 원활하게 하기 위한 목적을 달성하기 어려움이 예상되는 것은 분명해 보인다[4]. 이에 따라 본 연구는 농인 중심의 입장에서 한국어를 이해하기 위하여 한국어 텍스트를 한국수어로 번역하는 과정에서 필수적으로 요구되어지는 데이터 전처리에 관한 연구를 진행하였다. 특히, 한국어에서 한국수어로 번역하는 과정에서 텍스트를 중심으로 하는 한국어/한국수어 기계번역 연구를 수행하며 번역 성능의 고도화를 위해 필수적으로 요구되는 텍스트 데이터 전처리 기법에 대한 연구를 진행하였고, 이와 연계되어 한국어를 한국수어로 번역하는 과정 중 최종적으로 아바타 애니메이션을 통해 한국수어를 실제로 표현하기 위해 데이터 수집 과정에서 필수적으로 요구되는 한국수어 텍스트 데이터 스크립트 작성 기준을 설계하였다. 또 해당 설계 기준에 따라 데이터를 직접 수집하고 기계학습을 위한 자연어 전처리 기법에 대한 알고리즘을 구현하고 실제로 프로그램을 개발하여 한국수어 데이터 전처리 과정에 적용하였다. 본 연구는 한국수어를 최종적인 아바타 애니메이션으로 표현하여 청인과 농인 상호간의 의사소통을 원활하게 진행할 수 있는 텍스트 데이터 제작에 관한 기준을 농인 중심적으로 정립하고 수집된 데이터에 대한 실질적 전처리를 수행하여 한국수어 데이터의 정제 연구를 수행하였다는 것에 의의가 있다.

1-2 연구 목적 및 방법

본 연구의 목적은 한국어에서 한국수어로 단순 텍스트 번역이 아니라 한국어로 작성된 문장 단위의 성문 텍스트 데이터를 아바타 형식의 애니메이션으로 번역하여 효과적으로 표현하기 위한 전 단계로써 한국수어 문장 스크립트 제작 기준 및 텍스트 데이터 전처리 기법을 연구하는 것에 있다. 인공지능 분야에서 기계학습을 위한 파이프라인은 일반적으로 문제를 정의하고 데이터를 수집하며, 수집된 데이터를 분석 후 전처리를 진행한다. 이후 알고리즘을 적용하는 방식으로 파이프라인이 완성된다. 본 연구에서는 이러한 일반적인 파이프라인을 수어 번역과 표현의 특수성을 고려하여 데이터 수집과 전처리를 중심으로 구성된 파이프라인을 재구성하였다.

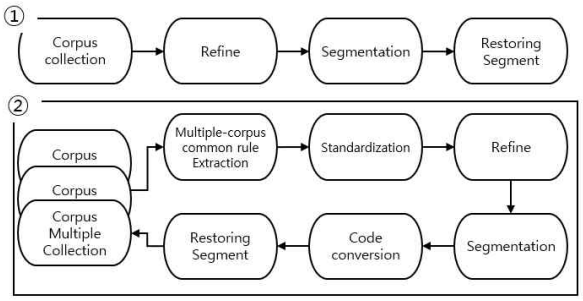


그림 1. 한국수어 전처리 파이프라인 비교도
Fig. 1. Comparison of Korean sign language preprocessing pipeline

본 연구에서는 그림 1의 ①과 같이 코퍼스를 단순하게 수집, 정제, 분절, 분절 복원의 4단계의 접근방식과는 다르게 한국수어 번역모델의 최종적인 성능을 농인에게 효과적으로 전달하기 위해서 대규모의 데이터 수집이 반드시 필요하다는 점을 고려하였다. 코퍼스 수집 단계에서 기계학습을 진행할 수 있을 만큼의 데이터를 확보할 수 있다면 기존 4단계 방식의 일반적인 텍스트 데이터 전처리 기법을 활용한 기계번역의 성능을 보장할 수 있지만, 한국수어 데이터와 같이 데이터의 절대 수량이 부족한 특수한 영역에서는 충분한 데이터 수집이 어렵다는 것이 일차적인 문제이고, 근본적으로 한국어를 한국수어로 번역하는 표준이 없기 때문에 수어번역가들이 자신만의 수어번역 방식에 따른 불규칙적인 원시 번역 데이터를 생성할 수 밖에 없다는 문제가 있어 왔다. 음성언어와 달

리 수어는 시간언어일 뿐만 아니라 문자로 표기하기에 상당한 제약이 따르므로 기록이 용이하지 않다. 수어 표기법들은 일반적으로 습득하는 데에 상당한 시간과 노력이 요구된다. 훈련된 여러 명의 연구자들이 표기 작업만을 매우 오랜 시간 동안 진행해야만 가능한 접근이다. 결국 수어 표기법을 상용화하기란 현실적으로 매우 어렵다[5]. 통계적 기계 번역은 다음의 문제점들을 야기한다. 첫째, 양 언어 표현 간에 의미적 맵핑 관계를 추론하는 구 단위 표현 정렬의 정확도를 떨어뜨린다. 둘째, 비 번역된 단어들이 다수 발생할 가능성이 높다. 셋째 번역 시 단어의 성질에 따라 달라지는 어순의 변화를 학습하는데 제약이 있다[6]. 이러한 문제점들은 결국 적은 양의 말뭉치에 출현하는 어휘의 종류가 제한적이며, 병렬 문장 쌍들에서 동시에 출현하는 이중 언어 표현 빈도가 낮고, 어순의 변화로 인하여서도 서로 상이한 단어들과 정렬될 가능성이 높아 잘못된 번역이 된다는 것이다. 이와 같은 문제점을 해결하기 위하여 두 가지 접근방식이 존재한다.

첫 번째 정석적인 방법은 먼저 한국수어번역 표준을 수립한 후 대규모 데이터를 확보하는 것이 있지만 현재 상황에서는 해결하기 어렵고 두 번째는 기존에 수집된 데이터를 종합하고 공통된 규칙으로 전처리하여 대규모 데이터셋화 하는 방식이다. 이에 따라 본 연구에서는 현재 적용이 가능한 두 번째 해결법으로 접근하여 ②의 파이프라인과 같이 형식이 다른 코퍼스를 다중으로 수집하고 공통의 규칙을 추출하여 표준화하며 수어 애니메이션 생성을 위한 코드 변환까지 할 수 있는 7단계의 파이프라인을 설계하였으며, 데이터 제작 기준의 설계와 전처리 방식을 기초단계부터 수립하는 방식으로 연구를 진행하였다. 기존의 다른 연구들은 수어 번역문과 수어 동작 표현의 연계를 고려하지 않은 단순 텍스트 번역으로 한국어 문장을 한국수어 문법 형식에 따라 정확하게 변환하는 것을 주요한 결과물로 하였으나, 해당 결과물로 실질적 수어 애니메이션을 어떻게 생성할 것 인가에 대한 연구는 다루지 않았다[7]-[9]. 농인들에게 수어문법에 의해 작성된 텍스트만을 전달하는 것은 실질적으로 한계가 있어 아무리 높은 수준의 번역 성능을 가진 모델을 개발하더라도 실효성이 거의 없었다고 할 수 있다. 이와 같은 방식으로 개발된 데이터들은 농인에게 내실화 및 고도화되지 못하고 실질적인 서비스나 제품으로 환원되지 못하고 있다[10].

본 연구과정에서 제안한 7단계 파이프라인은 번역된 수어 문장이 텍스트 형태에서 전처리 과정이 완료되는 것이 아니라 기존의 분절과 분절 복원 사이에 코드 변환 단계를 추가하여 각 수어 단어에 매칭 될 수 있는 수어 동작 코드를 탐색하고, 이를 변환하여 수어 애니메이션을 즉각적으로 생성할 수 있는 서비스에 적용하여 수어 번역문과 수어 동작 표현의 연계를 할 수 있다. 본 연구는 수어 데이터 성과물을 형성하는 것에 매몰되지 않고 한국어로 제작된 콘텐츠를 농인들이 이해할 수 있게 전달하기 위해서는 한국어가 아닌 한국수어의 형태로 제공해야 정보접근성이 높을 수 있다는 선행연구에 따라 콘텐츠에 대한 해설을 텍스트가 아닌 수어 애니메이션을 생성할 수 있는 최종적인 서비스단계까지 제공하기 위해 데이터 활용도를 최대한으로 향상시키는 것을 목적으로 하였다. 또한, 본 연구는 문화체육관광부 주관 한국콘텐츠진흥원의 정책지정 과제로 수행한 한국문화기술연구소의 “청각장애인을 위한 지능형 전시해설 문자/한국수어 변환 기술개발” 및 “청각장애인과 청인의 의사소통을 위한 인공지능 기반 수어 통번역 서비스 기술개발” 연구 과정에서 수집된 한국어-한국수어 문장 병렬 데이터와 국립국어원의 “2022년 한국어-한국수어 병렬 말뭉치 구축” 사업 등에서 이관 받은 데이터 총 147,518개의 문장을 기본 데이터를 대상으로 수행하였다.

다음 장에서 자연어처리 분야에서 기계학습을 위한 최근 이론에 대한 배경을 기술하고 연구를 수행하면서 데이터 그룹으로 특정한 데이터셋을 개별 분석하고 공통속성으로 특징되는 점들을 유형화하였다. 이후 기계번역을 위한 한국수어 데이터 스크립트를 제작하면서 요구되는 개선점을 도출하고 해당 데이터로 한국수어 3D 아바타 제어를 위한 한국수어 모션 텍스트 데이터 매칭 기호들을 공통적으로 전처리하기 위한 알고리즘을 개발하고 적용하는 방법으로 연구를 진행하였다.

II. 한국수어 전처리를 위한 선행 연구 및 관련 이론

2-1 기계번역을 위한 대형언어모델

연구 배경에서 기술하였듯이 현재 한국수어와 관련된 연구들이 영상처리 분야에 집중되어 있고 번역 성능의 기대치가 낮은 이유 중 하나는 높은 성능을 갖춘 대규모 인공지능 모델에 기반한 기계 번역기를 한국수어 텍스트 번역에 적용하기 어렵다는 이유에서 기인한다[11]. 일반적으로 대상언어와 목표언어 두 개의 언어를 번역하는 과정은 두 대상에 대한 텍스트를 컴퓨터가 이해할 수 있는 자연어 처리방식에 따라 정제한 이후 학습모델에 의해 기계번역을 수행하게 된다[12]. 이 과정 중 수많은 단계가 내포되어 있고 규칙기반 기계번역(Rule Based Machine Translation, RBMT)부터 통계기반 기계번역(Statisctcal Machine Translation, SMT) 그리고 최근의 인공신경망 기계번역(Neural Machine Translation)까지 다양한 기계번역 기법들이 있지만 결론적으로 원래의 대상언어에서 목표언어의 문법과 규칙성을 반영하여 치환하는 방식인 입력과 출력 두 단계로 구성되어 있다고 할 수 있다[13]. 자연어처리를 위한 기계학습 알고리즘의 고도화로 인하여 요즘에는 대상언어에 대한 레이블 없이 양 언어 관계에서 동일한 의미를 지니는 병렬 데이터셋을 수집하고 학습시켜 번역문장을 생성하는데 이는 종단 간(End-to-End) 방식의 특성을 갖는 기법이라고 하며 입력 값을 받는 인코더(Encoder)와 번역 값을 출력하는 디코더(Decoder)로 구성된다[14]. 해당 모델들은 최소 수백만개 이상의 광범위하고

방대한 양의 문장 데이터를 이용하며, 장시간의 학습 시간이 요구되는데 이러한 특징들 때문에 최근의 자연어처리를 위한 모델들이 대형언어모델(Large Language Models, LLM)로 불리게 된 것이다[15]. 해당 모델들의 성능 평가 보고서를 보면 인간이 번역하는 결과보다 우수한 성능 수준을 보여주며 언어 번역분야의 도전과제를 근사하게 정복했다고 평가받기도 하지만 많은 양의 훈련 데이터 세트를 확보하기 어려운 한국수어 연구와 같은 특수성이 있는 영역에서는 이와 같은 높은 성능을 실질적으로 기대하기 어렵다는 것을 다수의 연구자들이 동의하고 있으며 이러한 이유 때문에 실제로 완성도 높은 수어 번역 모델이 없는 것이다[16].

2-2 한국어와 한국수어에 대한 전처리와 임베딩

자연어 처리에서 새로운 방식의 딥 러닝과 고도의 자연어 처리 알고리즘을 개발하는 것보다 실질적으로 가장 중요하게 여겨지는 부분은 기계가 학습할 데이터를 전처리(Preprocessing) 하는 과정이다[17]. 자연어처리 분야의 연구자들 상당수가 이에 동의하지만 실제로는 짧은 연구 기간과 소수의 연구인력, 낮은 연구비 규모, 높은 인적자원의 소모 등의 연구 제약사항으로 인하여 데이터 전처리를 소홀히 하거나 기피하고 있는 것으로 사료된다[18]. 본 연구에서는 이러한 점을 개선하기 위하여 기계번역 알고리즘 개발 자체를 목적으로 하기 보다 데이터를 전처리를 정확하게 처리하며 자동화하는 과정에 집중하였다. 일반적인 텍스트 언어 데이터를 전처리하는 과정은 기본적인 특수문자 제거, 조사제거를 위한 형태소 분리와 띄어쓰기 분리 등의 다양한 전처리 기준에 따라 문장 데이터를 단어 형태의 토큰화로 치환시켜 최종적인 개별 글로스(gloss)로 도출하는 기계적 처리를 수행한다[19]. 한국어-한국수어 번역은 데이터 부족 문제로 인하여 대형언어모델에 기반한 기계학습을 진행할 수 없는 것이 통상적이기 때문에 일반적인 데이터 전처리 과정과는 다르게 한국수어의 특수성을 데이터 생성 초기 단계부터 고려해야 원활한 기계번역이 가능하다[10].

한국어와 한국수어의 언어 유형학상의 본질적인 차이점으로 한국어는 어근에 접사가 결합하여 단어의 기능이나 의미가 변화하는 형태의 교착어라는 점이고, 한국수어는 어형 변화나 접사가 없고 위치에 의하여 단어가 문장 속에서 가지는 여러 가지 관계가 결정되는 언어로써 단어가 변화하지 않고 단어의 순서로 문법적 기능을 나타내는 고립어적 특성을 가지고 있는 점에서 차이가 있다[20].

자연어처리 분야에서 임베딩(Embedding)은 사람이 쓰는 자연어를 기계가 이해할 수 있는 숫자의 나열인 벡터(Vector)로 바꾼 결과 혹은 그 일련의 과정 전체를 의미한다[21]. 기계번역을 위해 해결해야 하는 문제들은 단어와 문장 임베딩을 사전 학습하는 업스트림 태스크(Upstream task)와 다운스트림 태스크(Downstream task)로 구분할 수 있으며, 후자에서 구체적으로 처리 해야할 주요한 문제들은 품사 판별(Part-Of-Speech tagging, POS), 개체명 인식(Named Entity Recognition, NER), 의미역 분석(Semantic Role Labelling) 등이 있다[22]. 단어의 순서가 임의로 배치되더라도 의미가 크게 달라지지 않는 한국어 임베딩과 다르게 한국수어 임베딩 과정에서는 체언이나 부사, 어미에 연결되어 문법적 관계를 표시하거나 의미를 보조해주는 품사인 조사가 없어 어순을 제약해야만 의미해석이 달라지지 않기 때문에 품사 판별 태스크를 정확히 수행하여 순서를 준수해 생성하는 것이 중요하다[23].

표 1. 한국수어 기본문장의 고립어적 특성 예시
Table 1. Example of isolated linguistic characteristics of basic sentences in Korean sign language

No	Korean sentences	Korean sign language sentence
1-1	아내가 예쁜 딸을 낳았다.	아내 / 낳다 / 딸 / 예쁘다
1-2	아내가 낳았다 예쁜 딸을	아내 / 낳다 / 딸 / 예쁘다
1-3	예쁜 딸은 아내가 낳았다.	아내 / 낳다 / 딸 / 예쁘다
2-1	예쁜 아내가 딸을 낳았다.	아내 / 예쁘다 / 딸 / 낳다
2-2	딸을 낳은 아내는 예쁘다.	아내 / 예쁘다 / 딸 / 낳다

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

표 1의 구분 1 “아내가 예쁜 딸을 낳았다.” 라는 한국어 문장에서는 단어 순서가 서로 변화하여 교착되더라도 한국어에서는 같은 의미를 지니지만 한국수어 문장에서는 문장 구성이 고립적으로 동일한 순서를 지켜야 올바른 의미 전달이 될 수 있다. 구분 2에서 “예쁘다”라는 의미를 아내를 대상으로 수식하며 의미가 달라질 때도 이는 마찬가지 특성을 보인다. 구분 2-2 한국어를 “딸 / 낳다 / 아내 / 예쁘다” 같은 방식으로 순서를 지키지 않고 한국수어로 번역을 하게 되면 “딸이 예쁜 아내를 낳았다”와 같은 이상한 의미의 수어문장으로 번역되어 혼란을 발생시킬 수 있다.

데이터 전처리 과정에 텍스트 임베딩은 기계번역 모델의 성능을 향상 시킬 수 있는 요소지만 실무상에서는 기존의 한국어 임베딩 또한 영어와 같은 초대규모 데이터 규모에 비해 양적으로 데이터가 부족하여 그 효과가 크지 않다고 알려져 있다. 결론적으로 한국어보다도 훨씬 더 데이터 셋 규모가 작은 한국수어에 대해서는 임베딩 기법을 적용할 수 없었고, 실제로 천여개의 문장 단위의 데이터셋만이 존재하는 한국수어 텍스트 데이터에 대한 고립어적 특징을 반영한 전처리 기법도 찾아볼 수 없었다. 이에 따라 기존의 임베딩 기법과 단어 순서를 교체하는 방법으로는 한국수어 번역 모델의 성능을 향상 시킬 수 없는 것으로 판단하였다. 선행되는 문제를 해결하기 위해서는 무엇보다도 대규모 데이터셋을 확보하는 것이 중요한 과제이므로 본 연구에서는 기존에 산발된 한국수어 데이터셋을 규칙성 있게 통합하는 것으로 해결 방법을 제안하였다.

Ⅲ. 한국수어 데이터셋 분석 및 유형별 분류

인공지능 기반 기계번역 모델을 학습시키고 성능을 평가하는 과정에 있어서 필수적으로 지켜져야 하는 것은 공통의 형식을 갖는 데이터셋을 사용하여 학습시키고 테스트를 진행하며 검증을 했는지의 여부이다. 대형언어모델을 학습시킬 수 있을 만큼의 광범위하고 방대한 양의 병렬 데이터셋을 활용할 환경이 되지 못한다면, 같은 언어의 데이터셋을 사용하는 경우에도 각기 다른 패턴의 데이터셋을 사용하여 인공지능 모델을 기계학습시키는 것은 영어로 모델을 학습시킨 뒤 중국어로 번역 테스트를 진행하는 것과 다르지 않은 결과를 보일 것이다. 데이터 수집을 시작한 때부터 완료 시까지 정해진 규칙에 따라 데이터를 생성하고 수집하는 것이 최선의 방법이지만, 연구 프로세스상 다수의 연구자들이 단계적으로 연구를 참여하며 수집, 생성된 데이터가 다양한 패턴을 형성하는 것은 불가피한 결과이다. 이에 따라 본 장에서는 한국수어 번역모델을 개발하면서 수집하였던 병렬 데이터셋의 특징을 분석하여 유형별로 분류하고, 이를 전처리 알고리즘에 공통적으로 적용할 수 있는 방식에 관해 연구하였다.

현재 본 연구과정에서 대상으로 선정한 한국수어 텍스트 데이터는 “청각장애인을 위한 지능형 전시해설 문자/한국수어 변환 기술개발” 및 “청각장애인과 청인의 의사소통을 위한 인공지능 기반 수어 통번역 서비스 기술개발” 연구단계에서 수집된 한국어-한국수어 문장 병렬 데이터와 국립국어원의 “2022년 한국어-한국수어 병렬 말뭉치 구축” 사업 등에서 제공 받은 데이터로 총 147,518개의 문장이다.

표 2. 한국수어 텍스트 데이터 수집현황
Table 2. Collection status of Korean sign language text data

No	Dataset name	Quantity	Key Features	Agency
1	GKSL	13,047	data augmentation	Seoul National University .etc
2	2022 Korean-Korean sign language parallel a bunch of word	120,295	Medical Civil Service Administration Shopping and Tourism Area	National Institute of Korean Language
3-가	CTKSL-1	8,795	Cultural Area	KCTI
3-나	CTKSL-2	3,791	Self Normalization	
3-다	CTKSL-3	1,590	Common rules and standardization	
Total Sentence Quantity		147,518		

각각의 데이터셋은 수집된 한국어 문장을 한국수어로 번역하는 과정에서 한국수어문과 수어영상에 대한 수집을 목적으로 번역가와 수연가인 인간의 이해를 돕기 위해 제작되었다. 번역된 수어 문장을 게임엔진 기반 수어 애니메이션으로 자동 생성하기 위해서는 기계인 컴퓨터가 이해할 수 있는 방식으로 동일한 명령어를 사용해야 하므로 해당 데이터셋의 개별 표기 방식에 대한 유형별 분류를 선행하고 통합규칙을 정립하는 것이 필요하다.

3-1 GKSL(Gloss level Korean Sign Language)

GKSL은 Automatic Gloss-level Data Augmentation for Sign Language Translations 논문을 통해 발표된 한국수어 데이터셋이다. 수어 번역 모델링에서 데이터 부족 문제를 극복하기 위해 3,052개 원시 문장을 수집하고 텍스트 데이터 증강 기법인 공백 대체(Blank Replacement, BR), 구문 재해석(Sentence Paraphrasing, SP), 동의어 대체(Synonym Replacement, SR) 세 가지 기법을 사용하여 13,047개로 증강 시킨 데이터셋이다. 데이터 품질 평가를 위해 수어 전문가에게 검증을 진행하였지만 500개의 샘플 중 173개가 부정확하다는 평가를 받았다. 초기에 해당 데이터 셋으로 기계학습을 진행하며 테스트를 수행하였지만, 데이터의 정확도가 낮고 일관성이 부족한 패턴 때문에 최종적으로 본 연구결과에 적용하기 적합하지 않은 데이터셋이었다.

3-2 국립국어원 한수병렬 말뭉치 데이터

국립국어원은 “2022년 한국어-한국수어 병렬 말뭉치 구축”사업을 통해 생활 분야 의료와 관련된 15,158개의 문장과 민원행정에 대한 43,542개, 문화 분야 쇼핑 29,052개, 관광 32,543개까지 총 4개 영역에서 120,295개의 상당한 양의 한국어/한국수어 병렬 문장 데이터셋을 구축하였다.

표 3. 한국어-한국수어 구축 데이터 수량 [24]
Table 3. Data quantity of Korean-Korean sign language construction[24]

Classification		Medical care	Civil service admini-stration	Shopping	Tourism	Sum
Kor	Senten-ce	15,158	43.542	29,052	32,543	120,295
	Word	273,663	303,713	203,370	220,341	1,001,087
KSL	Senten-ce	15,158	43.542	29,052	32,543	120,295

한국어 어절 기준으로 백만개가 넘는 양의 데이터를 구축하였으며 한국수어 형태로 라벨링 작업을 진행하며 주석처리에 대한 세부 지침도 수립하였다. 해당 과정에서 한국어 문장 추가정보를 기입하고 있는 주요 지침 중 첫 번째 사항은 동적 숫자 표기법에 관한 것이다.

해당 데이터셋에서는 수어의 특성상 수치를 나타내는 표현을 n과 d로 구분하여 단순 숫자, 시, 시간, 날짜, 나이를 명확히 구분하여 별도의 표기법으로 한국어 문장에 표기하였다.

표 4. 국립국어원 데이터셋 동적 숫자 표기법 예시[24]
Table 4. Example of dynamic numeric notation for the data set of the National Institute of the Korean Language[24]

Notation		Class	Definition	Example of notation
n		Number	Applies to simple numbers, such as the number of objects	n:두개
d	Hour	Dynamic number	Apply to specific time expressions such as what time	d:시:일곱시
	Time		Applies to time zone expressions such as hours	d:시간:세시간
	Date		Applies to month and day expressions, such as what date and month	d:날짜:시월이십일
	Age		Apply to the expression of age	d:나이:아홉살

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

수에 대한 표현을 일반숫자(n:)와 동적숫자(d:)로 구분하고 시(d:시), 시간(d:시간), 날짜(d:날짜), 나이(d:나이)로 세분화하여 어떤 개념과 관련된 숫자의 표현인가를 인간이 이해하기 위한 기준으로 정한 것은 기계번역과 애니메이션 표현을 위한 방식에서는 불필요하여 후술할 CTKSL-2의 숫자 표현과 같은 방식이 더 적합하다.

표 5. 국립국어원 데이터셋 동음이의어 표기법 예시[24]
Table 5. Example of the Korean Language Institute's dataset homophonic notation[24]

Word	Class	Definition	Example of notation
배	body	means the belly of the body	배(body)가 아프다
	Fruit	It means fruit	배(과일)는 싫다
	배수	두배, 세배 등의 배를 뜻함	배(배수) 이다
	교통	보트, 등의 배를 뜻함	배(교통) 타다
다리	신체	신체의 다리를 뜻함	다리(신체) 아프다
	교량	시설물 다리를 뜻함	다리(교량) 생기다

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

한국어를 한국수어로 번역하며 오역이 발생할 경우를 대비하여 다양한 동음이의어를 표기하기 위한 구분으로는 문장 뒤 괄호표시 “()”에 보충의미 삽입을 통해 의미 전달에 혼란이 없도록 표기법을 정의하였다.

생산적 수어는 일반적인 수어 단어 표현에서 벗어나 수어의 도상성을 활용하여 특정 상황 및 문장의 맥락 속에서 사용되는 표현을 의미한다[24].

표 6은 국립국어원에서 제작한 생산적 수어의 예시이다. 해당 데이터셋에 대한 결과보고서에서는 생산적 수어와 관련하여 한국어 문장에 대응하여 한국수어로 번역할 때 상황묘사와 형태 변화, 정도 묘사를 위해 생산적 수어에 타입 입력만 진행하였다고 밝혔으며 현재 AI(Artificial Intelligence)의 학습 능력이 생산적 수어를 인지하기에 한계가 있다고 기술하였다. 시각 언어라는 특수성에서 비롯된 시각적 표현과 공간을 활용하여 특정 상황을 해설하기 위해 마치 그림처럼 표현하는 수어의 도상성 표현은 사용 빈도가 일회성에 그치고 한국어 문장과 한국수어 글로스 형태가 표 6의 한국수어 예시 문장과 같이 상이 하여 인공지능 기계학습에 적절하지 않기 때문에 본 연구에서는 저자원언어(Low-Resource Language) 환경을 극복하기 위한 저빈도규제(Low Frequency Regulation) 방법론을 적용하여 생산적 수어와 비수지 신호와 관련된 데이터를 규제하였다.

표 6. 국립국어원의 생산적 수어 예시[24]
Table 6. Examples of productive sign language at the National Institute of the Korean Language[24]

Korean sentences	Korean sign language sentence
차가 굽은 길을 달리다.	자동차1(오른쪽방향 돌면서)
바람이 강하게 분다.	바람(거세게)
큰 상자	상자(큰 상자)

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

3-3 한국문화기술연구소 한수병렬 데이터

1) CTKSL-1

한국문화기술연구소에서 2020년부터 2022년까지 수행했던 “청각장애인을 위한 지능형 전시해설 문자/한국수어 변환 기술개발” 연구과제에서는 한국어-수어 1대1 스크립트 데이터셋 제작 방식에 관한 연구를 진행하며 인공지능 알고리즘 적용에 유효한 데이터셋 형식을 정의하였다. 본 논문에서는 해당 데이터셋을 CTKSL-1(culture technology korea sign language Level 1)로 명명하였다. CTKSL-1은 수어 전문가들이 제안했던 수어 문법에 기초한 방식을 인공지능 입력과 출력 알고리즘에 맞게 재가공하는 문법을 사용하였다. 연구결과를 통해 도출한 데이터셋은 텍스트 형태의 결과물로 해당 단어의 모션 데이터만 존재한다면 아바타로 수어 문장을 생성할 수 있는 방식으로 실제 인공지능 알고리즘에 적용할 수 있는 효과성을 입증한 데이터셋 형식이다[20]. 현재의 연구과제에서도 해당 데이터와 1대1 매칭 알고리즘을 고도화하는 방향으로 연구를 진행하고 있으며 데이터 구성은 다음과 같다. 공공시설에 기반한 전시해설을 한국수어로 서비스하기 위한 목적으로 문화 영역에 특정하여 8,795개의 한국어 문장을 수집하였으며, 전문 수어 번역가들과 연구자들이 이에 대응하는 5,581개의 문장을 공동으로 작업하며 한국수어문장을 생성하였다. 해당 문장 중 223개의 문장은 게임엔진에서 구동될 수 있는 모션캡쳐 FBX(Filmbox) 데이터와 매칭되어 3D 수어 아바타 애니메이션을 자동으로 생성하는 프로그램 개발을 통해 국립광주박물관과 배재학당 역사박물관에서 농인들을 대상으로 하는 전시 실증 서비스에 사용되었다.

표 7. CTKSL-1 데이터셋 예시문[25]
Table 7. Example sentences for CTKSL-1 dataset[25]

No	Korean sentences	Korean sign language sentence
1	'분청사기'는 옛 기록에 없는 이름입니다.	분청사기 / 이것 / 옛 / 기록 / 이름 / 없다
2	아펜젤러는 배재학장을 설립하면서 통역관 양성학교의 통념을 타파했다.	아펜젤러 / 남자 / (왼손) / 엄지형 / (오른손) / 지칭 / 배재 학교 / 설립 / 동시에 / 통역 / 전문 / 교육 / 학교 / 법 / 생각 / (두손)버리다
3	길상무늬 청화백자가 유행하였습니다.	길상 / 무늬/ 청화백자 / 청화백자 / (지화) / 유행

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

CTKSL-1 데이터셋의 가장 기본적인 형식은 '단어' + '단어 문법정보'로 구성되어 있다. 수어 단어에 특정한 문법 정보가 추가되어야 할 필요성이 있는 경우에는 해당 단어 뒤에 문법 정보를 괄호 "()" 속에 추가 입력하였다. 표7 2번 예시문에서 "남자", "(왼손)", "엄지형", "(오른손)" 은 남자 수형은 왼손으로 엄지형 수어는 오른손으로 표현하라는 지시이다.

예시문 1번과 2번 한국수어 문장에서 볼드체로 진하게 표시된 부분은 지화처리를 하기 위한 부분을 인간이 인지하기 위한 목적에서 사용되며 아바타 애니메이션 구현상에서는 한국수어 사전에 모션데이터가 없으면 모두 지화로 구현하는 알고리즘을 적용하는 방식으로 작동한다. 그러나 해당 단어를 지화로 표현하는 경우와 지화가 아닌 수어 단어로 처리하는 경우를 구별하는 경우를 대비하여 강제로 해당 단어를 지화처리하는 문법 정보를 추가하여 예시문 3번 "청화백자 / 청화백자 / (지화)" 처럼 "(지화)"를 추가해주는 방식을 고안하였다. 해당 방식은 중복되는 단어가 불필요하게 반복되어 의미의 혼란을 주고 자동기계번역 방식에서 출현 빈도가 낮은 희소성 있는 문장들에 대한 학습은 올바르게 이루어지지 않기 때문에 결국에 컴퓨터가 이를 구분하여 번역할 수 없는 문제가 발생한다. 한국수어 문장을 코드로 변환하는 개발자의 입장에서는 한국수어 번역 문장을 제작한 사람이 "(지화)"라는 지시어가 앞과 뒤에 있는 단어 중 어느 단어에 대해 지화로 표현해야 한다고 기록하였는지를 직관적으로 해석하기 어려워 문맥에 따라 해석을 해야 하고 이에 따라 알고리즘 분기점 설정을 기계적으로 진행하기 어려운 문제가 있었다.

따라서 현재의 CTKSL-3의 데이터셋에서는 번역된 한국수어 문장의 전처리 과정에서 2차적으로 문장을 다시 정제하여 애초에 단어사전에 모션데이터가 존재하는 경우에는 수어 단어로 표현하고 부재하거나 지화처리를 필요로 하게 되는 경우에는 해당 단어 앞에 "(지문자)" 명령어를 삽입시키는 방식으로 인간과 기계 모두 직관적인 방식으로 인지하고 처리할 수 있는 알고리즘으로 개선하였다.

당시 연구자들은 수어의 독특한 특성 때문에 발생하는 특별한 단어들을 "분류사"로 정의하였다. 예를 들어 "바르다"의 뜻을 가지는 수어 동작은 도자기에 색을 바를 때와 건물 벽면에 페인트를 바를 때의 수형이 달라지므로 "(건물에)바르다", "(도자기에)바르다"와 같이 분리하여 기록하고 전자는 가상의 건물 벽면 모양을 가상으로 놓고 크게 바르는 수형으로 표현하며, 후자는 가상의 도자기 모양을 두고 바르는 수형으로 별개의 취급을 하였다. 그러나 해당 방식은 바르는 객체가 변화함에 따라 수형을 계속적으로 수집해야 하는 필요가 발생하고 한국수어 문장으로 번역할 때 일회성으로 사용되기 때문에 인공지능을 활용한 기계번역 방식에 적합하지 않아 개선이 필요하게 되었다.

2) CTKSL-2

CTKSL-2(culture technology korea sign language Level 2)는 번역기 기반 수어 캐릭터 구동 범위를 요소별로 정의하며 CTKSL-1의 개선점을 매뉴얼화 하고 이를 적용하여 추가적으로 수집한 3,791개의 데이터를 정규화한 데이터셋이다. 기존의 데이터셋에서 구별되는 주요한 특징으로 각 지화와 영문 단어 앞에 "(지화)", "(영문)" 이라는 토큰을 사용하였으며 영어문자는 모두 대문자로 통일하였다.

표 8. CTKSL-2 숫자 유형별 데이터셋 예시
Table 8. Example of CTKSL-2 number type dataset

Class	Raw data	Annotation data
Cheirology	Appenzeller	(Cheirology)Appenzeller
English	QR	(English)QR
Number-1 (large number)	1987	1000+900+80+7
Number-2 (small number)	198.7	(Left Hand) 100+90+8 (Right Hand) 이것 + 7
Number-3 (natural number)	02-123-123	0+2+1+2+3+1+2+3
Number-4 (Number of units)	12Century	12 + Century
Number-5(Date)	9 Month 5 Day	9 Month + 5
Number-6(Year)	1987 year	1987 + year
Number-7(Period)	For 100Year	100 + Period

자연스러운 수어 아바타 애니메이션 생성을 위해 숫자 단위 표현에 대한 기준도 케이스별로 분류하여 명확한 규칙을 수립하였다. 기본적으로 독립수형이 존재하는 1부터 20까지 숫자의 모션데이터를 개별 수집하고 조합형으로 표현할 수 있는 이후의 수는 각 단위에 맞춰서 10만 단위까지 모두 수집하였다. 예를 들어 "Number-1(large number)" 유형의 1987과 같은 큰 단위의 숫자를 번역할 때 "1000+ 900+ 80+ 7"과 "1+ 9+ 8+ 7" 두 경우를 수어로 표현하는 것을 비교하였을 때 농인들은 원래의 의도대로 전자의 표현 방식이 가장 좋은 표현이라고 평가하였으며 후자의 표현 방식은 자릿수의 개념이 사라져 "1, 9, 8, 7" 이라는 각각의 숫자로 잘못 이해하였다. 이에 따라 전화번호와 같이 정확한 전달이 필요한

“Number-3(natural number)” 유형의 고유 숫자는 하나씩 끊어서 표현하였다. “Number-2(small number)” 유형에서는 소수점을 표현하기 위한 유형으로 소수점 앞자리는 왼손으로 오른쪽은 오른손으로 직관적으로 표현한다. 소수점은 “이것” 이라는 단어와 동일한 수형으로 표현하기 때문에 “이것”을 숫자 앞에 삽입하는 방식으로 소수점을 표현하였다.

“Number-4(Number of units)” 유형은 단위가 필요한 숫자로 처음에는 “12(Left Hand) + Century (Right Hand)”과 같이 양손을 사용해서 표현하였으나, 우세손인 하나의 손만을 사용해 왼쪽에서 오른쪽으로 이동하며 순차적으로 독립적인 수어처럼 표현하는 것의 이해도가 더 높다는 전문가의 의견에 따라 이를 규칙에 반영하였다. “Number-5(Date)” 유형에서 1월부터 12월까지는 독립 수형 데이터를 기본 수집하는 것을 전제로 한다. “9월 5일”처럼 월과 일을 동시에 번역할 때 일의 단위는 별도로 표현하지 않고 “9월 5”와 같이 번역하며 “Number-3(natural number)” 유형과 마찬가지로 우세손을 이용하여 한 손만으로 표현한다.

연도를 나타내는 “년”은 “때(year)”로 의미 전달성을 높이기 위해 치환하고 “1987year”을 “1987 + year”로 표현한다. 기간을 나타내는 “년”은 원래의 규칙대로 “년(Period)”으로 적용하여 “For 100year”을, “100 + Period”으로 번역한다.

표 9. CTKSL-2 데이터셋 예시문
Table 9. Example sentences for CTKSL-2 dataset

Class		Sentence
1	Kor	중국에 이어 한국과 베트남이 자기를 만드는데 성공하였고, 그 다음으로 일본이 만들었습니다.
	KSL	(공간) 중국 다음 (강하게) 한국 또 베트남 흙 그릇 만들다 (강하게) 성공 (강조) (휴지형) 다음 (강하게) 일본 만들다 (강하게) 합격 (바형)
2	Kor	광주·전남지역에서는 임진왜란 해전 당시 사용된 것으로 보이는 화포와 조총, 창과 검 등의 무기류가 많이 발견되었습니다.
	KSL	광주 전남 지역 일본 바다 전쟁 그 때 화포 (휴지) 발포 (휴지) 사격 장대투겁(창) 칼 2지로 검을 휘둘리는 동작 여러가지 발견 발견 발견

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

상기 숫자 유형별 기준은 번역 스크립트를 생성할 때 숫자 원문자 아라비아 형태로 기입하고 지화와 영어 유형을 스크립트 상에 표현할 때 처럼 공통적으로 수 표현 접두에 “(Number)” 키워드를 삽입한다. 한국수어에는 단순 수를 세는 용도가 아닌 연도 등의 서수형 숫자를 수어로 표현할 때 “1, 2, 3, 4, 10, 11, 12, 13, 14”와 같이 손등을 보여야 하는 표현이 있으므로 이를 고려하였다. 위 서수형 숫자가 사용되는 “7 Month 2 Day” 같이 특정 일자를 수어로 표현할 때는 Number-5(Date) 방식에 따라 “7 Month + 2”가 아닌 “7 Month + 2(손등)” 과 같이 표현해야 하기 때문에 기계번역을 진행할 때 예외적인 규칙처럼 번역해야 하므로 현재 단계에서는 데이터 표준화를 위해 적용하지 아니하는 것에 이점이 있다.

기타사항으로 “세기, 킬로미터, 미터” 등과 같은 단위 표현은 “C, KM, M”과 같이 전부 영문으로 통일하였다. 단위 또한 영문으로 표현되므로 영문 표현에서는 상기에서 정했던 “(영문)” 토큰을 전두에 삽입하여 단위 표현과 충돌을 방지하였다. 동음이의어의 경우 데이터를 생성할 당시부터 동형의 형태를 지니지 않도록 문장 뒤 의미 부연설명을 위한 괄호 형태 토큰을 추가하여 구분하는 방식을 적용하였다.

3) CTKSL-3

CTKSL-3(culture technology korea sign language Level 3)는 앞에서 언급했던 데이터셋들의 공통적 특징을 추출하고 수어 전문가가 수어문장을 생성할 때 인간의 실수를 줄이고 기계학습용 데이터의 효율을 높여 한국수어 번역모델에 적용하기 위한 목적에서 단순화된 수어문 작성규칙을 적용한 데이터셋이다.

앞서 생성하고 수집했던 수어의 도상성 문제를 해결하기 위하여 정의한 단순화 규칙의 주요한 특징은 도상적 표현으로 표현되는 단어는 최대한 일반적으로 통용되는 단어로 표기하고 복합명사는 단순명사로 분리하여 표현하였다. 한자어의 경우 한자의 의미별로 구분하여 최소한의 일반적인 단어로 표기하였다. 즉, 복잡한 의미의 복합어를 최대한 단순화하는 규칙이다.

표 10. CTKSL-3 단순화 규칙 적용의 예시
Table 10. Example of CTKSL-3 simplification rule application

Original sentence	Rule Transformation Sentence
슈트를 입는 모습(도상성)	정장/입다
금거북이	금/거북이
청소행정과	청소/행정/기관
일손을 더했다	도움/주다
규약, 규율, 원칙, 조항, 규정	규칙

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

다의어의 경우 앞선 데이터들과 마찬가지로 뜻을 명확히 하기 위하여 괄호안에 “배(과일)”, “배(신체)”, “배(운송수단)”과 같은 형태로 보조의미를 삽입한다. 의성어나 의태어는 맥락에 맞는 일반적인 단어를 사용하거나 배제하는 규칙을 적용하였다.

표 11은 CTKSL-3의 단순화 및 명료화 규칙을 적용한 예시문이다. CTKSL-2과 비교하였을 때 지시문의 수가 거의 없으며 비교적 쉬운 단어로 표현되어 있는 것을 확인 할 수 있다. 이는 수어 아바타 애니메이션을 생성하기 위한 요구사항과 이를 처리하기 위한 전처리 기법을 적용하기 위함이다.

앞에서 언급하였던 것처럼 도상성 표현을 수어 애니메이션으로 처리하기 위해서 기존에는 도상성 표현 전체에 대한 모션캡처를 진행하여야 하지만 CTKSL-3에서는 이미 기존에 가지고 있는 모션캡처 데이터와 변환이 가능하도록 최대한 의미가 통할 수 있도록 처리하였다.

표 11. CTKSL-3 데이터셋 예시문
Table 11. Example sentences for CTKSL-3 dataset

Class		Sentence
1	Kor	두 마을 사이에 자리한 아름다운 사원 사원 이름을 따서 이 지역을 왓츠베이 안덴 마을이라 하죠.
	KSL	마을/마을/가르다/중간/아름답다/절/있다/절/아름답다/유명/때문에/마을/이름/절/이름/활용하다/만들다/(지화)왓스베이안덴/마을
2	Kor	스승에게서 제자로 중앙아시아의 전통 문화를 전승하는 방식은 예전 방식이 많이 남아 있습니다.
	KSL	선생님/제자/중앙/아시아/문화/가르치다/모습/지칭/옛날/방법/스타일

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

표 12와 같은 수어 애니메이션 생성 요구사항에 따른 전처리는 단순한 규칙을 가지고 있지만 수어로 표현하기 어려운 문장들을 기존 데이터베이스에 저장 되어있는 쉬운 형태의 수어 모션 단어로 치환하여 활용할 수 있게 되기 때문에 게임엔진 기반의 프로그램에서 수어 애니메이션으로 즉시 변환이 가능하다는 이점이 있다는 발견점을 도출하였다.

표 12. 수어 애니메이션 생성 요구사항에 대한 전처리 기법 규칙
Table 12. Preprocessing technique rules for sign language animation generation requirements

Type	Original sentence	Pre-processing Transformation Sentence
Iconography	물 아래로 손을 넣어 조종하는 모습	물/아래/손/넣다/조종하다/모습
Polysemy	배	배(과일), 배(신체), 배(운송수단)
Similar word	규약, 규율, 원칙, 조항, 규정	규칙
figure of speech	일손을 더 했다.	도움/주다
Onomatopoeia	둥둥	떠다니다
Number	1박 2일, 만원	1박/2일, 10000
Part of speech	청소하다, 노래하다	청소, 노래
English	aBc	(지화)ABC
Sign	!, ?	삭제

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

다음 장에서는 먼저 본 연구에서 발견한 요인들로 데이터 전처리 알고리즘을 설정하고 연구 내용을 토대로 데이터 전처리를 순차적으로 진행한 실증과정을 서술하였다.

Ⅳ. 한국수어 데이터 기준 및 전처리 알고리즘

데이터 수집과정에서 한국어를 한국수어 문장으로 번역하는 수어번역가들의 작문 스타일이 개인별로 다르기 때문에 이를 공통적으로 처리하기 위해 문장을 개별단위로 최대한 분리하고 유사적으로 표현한 단어를 최대한 수어로 표현하기 위해서는 우리가 가지고 있는 한국수어 사전의 단어 데이터와 매칭하는 것이 가장 중요하였다. 앞서 정의한 규칙에 기반하여 가장 기본적이고 일반적인 단어를 사용하며 핵심 의미를 최대한 단순하게 정하고 띄어쓰기 대신 “/” 형태의 통일된 구분기호로 문장을 선제적으로 분리하였다. 이를 바탕으로 설계한 한국수어 데이터 전처리 알고리즘은 다음과 같다.

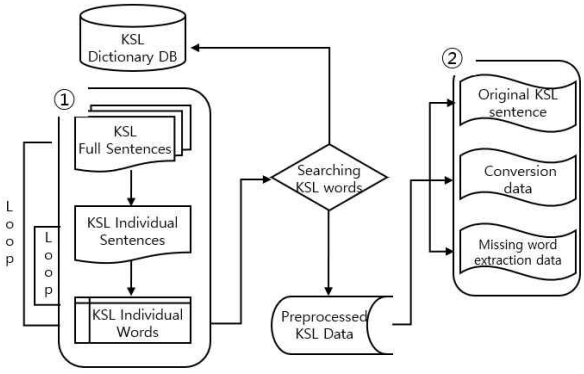


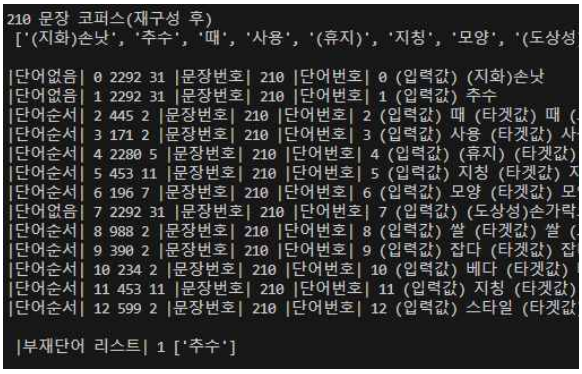
그림 2. 한국수어 데이터 전처리 알고리즘 구조
Fig. 2. Data preprocessing algorithm structure of Korean sign language

참고로 해당 알고리즘에서는 전처리 대상이 되는 한국수어 데이터는 프로그램 사용이 익숙하지 않은 수어 번역가가 손쉽게 활용할 수 있도록 하나의 개별 EXCEL 파일로 통합하여 처리할 수 있도록 구현하였다.

그림 2 ①모듈에서 번역이 필요한 전체 한국어 문장과 번역된 한국수어 문장이 포함된 하나의 EXCEL 통합 파일을 시스템에 업로드하면 알고리즘에서 문장 단위의 한국수어 문장으로 자동으로 분리한 뒤 하나의 개별 단어 단위로 글로스화 한다. 여기서 글로스화란 구분기호로 분리된 문장에서 각각의 단어가 별개의 의미를 형성하는 과정이다. 이렇게 토큰화된 개별 단어는 기존의 한국수어 단어사전 데이터베이스와 연동되어 순차적으로 표제어 또는 동의어에 존재 여부를 판별하게 되고 수어단어가 수어단어집에 존재한다면 최종적으로 수어단어의 고유번호를 매칭하여 수어 전처리 데이터로 통합한다.

그림 3은 한국수어 데이터 전처리 알고리즘 실행 결과의 일부이다. 210번 문장의 전처리 과정은 다음과 같다. 원래 문장 전체인 “추수에 사용되는 손낫은 둥근 칼 모양의 날이 바깥쪽을 향해 있어 손으로 별을 잡고 당겨가며 잘라내는데요.”와 번역 문장인 “(지화)손낫/추수/때/사용/(휴지)/지칭/모양/

(도상성)손가락으로 손낫 모양을 표현한다/쌀/잡다/베다/지칭/스타일"이 포함된 파일을 프로그램에 업로드하면 알고리즘은 해당 문장을 단어별로 분석하여 단어사전에 해당 단어가 존재하는지 여부를 일차적으로 판별하고 문장 내 단어 순서를 고정시켜 리스트화 한다. 수어단어는 한국어와 다르게 문맥상 동음이의어를 구별하기 어려우므로 전처리 알고리즘의 두 번째 단계에서는 문장 의미에 적합한 수어 모션단어를 정확하게 판별하기 위하여 해당 단어를 단어 데이터베이스 몇 번째 항목에서 찾아왔는지를 표시하는 정보를 추적한다. 예를 들어 단어순서 10의 "베다"의 경우 "받치다"의 의미와 "자르다, 가르다"의 두 가지 의미가 있는데 후자의 수어 모션을 애니메이션으로 표현하기 위해 단어 데이터베이스 234(베다)의 2번(자르다) 데이터로 모션을 매칭해야 한다는 의미이다.



*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

그림 3. 한국수어 데이터 전처리 알고리즘 실행 결과
Fig. 3. Results after the application of the algorithm of Korean sign language data preprocessing

②모듈에서는 위 프로그램에서 실행된 결과를 최종적으로 하나의 통합된 파일로 세분화하여 "원천 수어문장", "고유번호 변환 데이터(스크립트)", "부재단어 추출데이터" 세 가지 형태로 출력한다.

표 13. 한국수어 데이터 전처리 출력 결과
Table 13. Data preprocessing output results of Korean sign language

Category	Result
word_origin	(Cheirology)손낫 | 추수 | 때 | 사용 | …
DB	× | × | 445,2 | 171,2 | …
word_no	(Cheirology)손낫 | 추수 | 538 | 264 | …
nonWord_list	추수

*Categories are written in Korean to accurately convey the contents that were originally composed in Korean.

상기 한국수어 데이터 전처리 프로그램 개발 이전에는 위와 같은 데이터 전처리 작업을 위해 연구자가 문장에 속한 단어를 분리하고 해당 단어가 수어 단어집에 존재 여부를 판별하기 위해 각각의 단어를 하나씩 검색하는 수작업을 통해 진행했었다. 그 결과 많은 인적자원과 시간이 소모되었지만, 본 연구과정에서 개발된 자동화된 데이터 전처리 프로그램 사용으로 인하여 수 분내에 백만 어절 이상의 데이터를 정확하게 처리할 수 있는 비약적인 발전을 할 수 있었다. 해당 프로그램을 사용하여 최종적으로 수집했던 147,518개 문장 규모에 대한 전처리를 진행하여 기계학습에 적용할 수 있는 데이터 셋을 제작 하였으며, 분절 후 복원된 데이터와 변환된 코드가 일대일로 매칭되는 비율을 계산하는 방식으로 전처리 된 데이터가 잘 구성되었는지 평가를 진행하였다.

표 14. 전처리 데이터 구성 테스트 평가 주요결과
Table 14. Key results of test evaluation of preprocess data configuration

No	Test Sentence	Number of segment words	Non-convertible code	Accuracy rate(%)
1	32,543	210,237	79,424	62.2
2	1,000	5,133	718	86.0
3	43,542	285,260	58,788	79.3
4	3,000	20,994	2,229	89.3
		…		
30	33,542	221,880	25,424	88.5

테스트는 문장 개수를 임의로 정하여 알고리즘을 개선해 나가는 방식으로 30회 이상 순차적으로 비교평가를 진행하였다. 최초 테스트에서 정확률은 62.2%를 기록하였으나, 마지막 30차시 평가에서는 33,542개 문장 내 변환 불가 코드 25,424개를 제외한 분절 단어 221,880개를 전처리하며 88.5%의 높은 정확도를 기록하였다. 코드 변환에 실패한 단어는 현재 데이터베이스에서 존재하지 않는 새로운 단어이기 때문에 해당 코드가 없어 변환이 실패하는 것이며 새롭게 필요한 단어를 모션캡쳐하고 코드를 데이터베이스에 추가 하는 과정을 거쳐 변환 정확도를 높일 수 있다. 최종적으로 테스트 전체 평균 87.6% 정확률로 분절된 단어를 코드로 변환할 수 있었다.

V. 결 론

본 연구에서는 농인 중심의 한국어 의미전달을 위한 한국수어 스크립트 제작 및 텍스트 데이터 전처리에 대한 연구를 수행하며 수집한 한국수어 다중 코퍼스 문장에 대한 공통적 특징을 분류하고, 이를 바탕으로 한국수어 데이터를 전처리 할 수 있는 프로그램을 개발하였다. 한국어를 한국수어로 번역하는 과정에서 최신의 인공지능 기반 기계번역 알고리즘을 적용하기 위해서 대규모의 데이터셋이 필수적으로 요구되지만, 한국어를 한국수어로 번역하는 것과 같이 특수한 영역을 갖는 번역모델 개발과정에서 사용되는 데이터셋의 문제점 중

첫 번째는 학습시킬 수 있는 절대적 양이 부족하다는 것이고 두 번째는 한국수어문장 원천 데이터를 생성하는 수어번역가들의 번역문장의 문법과 형태가 일관되지 못하여 다양한 패턴의 데이터를 생성했다는 점이다. 결론적으로 대표적인 이 두 가지 문제점으로 인하여 지금까지의 인공지능 기반 한국수어 기계번역모델의 성능이 실효성 있는 결과를 도출하지 못하는 것을 확인하였다.

본 연구에서는 상기 연구결과에 기반 한 한국수어 데이터 전처리 알고리즘 개발을 통해 백만 어절 단위의 데이터를 전처리하여, 기계학습에 사용할 수 있고 게임엔진 기반의 한국수어 애니메이션 생성 프로그램에 바로 적용할 수 있는 한국수어 데이터셋을 제작하였다. 해당 데이터셋만으로 아직까지 완벽한 수어 동작을 생성할 수 있는 것은 아니지만, 테스트 결과 수어 애니메이션을 생성할 수 있는 코드 변환 정확률은 초기 평가 62.2%에서 마지막 30회차 평가에서는 88.5%로 약 3만개 문장에 대해 정확률 26.3%의 증가분을 기록하였다는 것에 의의가 있다.

향후 연구에서는 한국어에 기반 한 수어 표현인 수지한국어(청식수어)가 아닌 농인의 입장에서 수어의 의미 전달성을 보다 높이기 위한 진정한 한국수어(농식수어)를 표현하기 위한 방안으로 도상성 표현이 포함된 생산적 수어와 비수지 신호를 표현하기 위한 규정이 포함된 기준을 수립하는 연구가 필요하다. 최종적으로 이를 표준화하여 난립해 있는 데이터들을 정규화하고 통일성 있는 수어 빅데이터로 고도화한다면 다른 기계번역 연구 분야에서처럼 거대언어모델을 활용한 고성능의 수어번역과 자연스러운 고품질의 수어 애니메이션 생성을 기대할 수 있다.

감사의 글

본 연구는 2024년도 문화체육관광부의 지원에 의하여 한국문화기술연구소에서 이루어진 연구로서, 관계부처에 감사드립니다.

참고문헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ... and I. Polosukhin, “Attention is All You Need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach: CA, pp. 6000-6010, December 2017. https://doi.org/10.48550/arXiv.1706.03762

[2] H. Seong and H. Cho, “Three-Dimensional Convolutional Vision Transformer for Sign Language Translation,” *The Transactions of the Korea Information Processing Society,* Vol. 13, No. 3, pp. 140-147, March 2024. https://doi.org/10.3745/TKIPS.2024.13.3.140

[3] M.-S. Kim, M.-H. Yoo, J.-H. Jang, J.-Y. Choi, G.-E. Na, M.-S. Kim, and J.-J. Na, “Implementation of Deep Learning-Based Sign Language Acquisition Online Platform,” *Journal of Digital Contents Society,* Vol. 23, No. 11, pp. 2147-2157, November 2022. https://doi.org/10.9728/dcs.2022.23.11.2147

[4] J. H. Choi, H. Lee, and C. H. Ahn, “Study on Korean-Korean Sign language Translation Technology for Avatar Sign language Service,” in *Proceedings of the Korean Institute of Broadcast and Media Engineers Summer Conference*, Online, pp.459-460, July 2020.

[5] J.-W. Lee, “A Study on the Forms and Characteristics of Korean Sign Language Translation according to Historical Changes,” *Journal of the Korea Contents Association*, Vol. 21, No. 5, pp. 508-524, May 2021. https://doi.org/10.5392/JKCA.2021.21.05.508

[6] H. Park, J.-H. Kim, and J. C. Park, “Addressing Low-Resource Problems in Statistical Machine Translation of Manual Signals in Sign Language,” *Journal of KIISE,* Vol. 44, No. 2, pp. 163-170, February 2017. https://doi.org/10.5626/JOK.2017.44.2.163

[7] H. Sim, H. Sung, S. Lee, and H. Cho, “Sign Language Dataset Built from S. Korean Government Briefing on COVID-19,” *KIPS Transactions on Software and Data Engineering*, Vol. 11, No. 8, pp. 325-330, August 2022. https://doi.org/10.3745/KTSDE.2022.11.8.325

[8] Y. J. Kim, S. B. Hyun, and C. J. Park, “Translation Program for Converting Korean Text into Sign Language Text,” in *Proceedings of the Korean Association of Computer Education Summer Conference*, Online, pp. 107-110, August 2021.

[9] C. An, E. Han, D. Noh, O. Kwon, S. Lee, and H. Han, “Building Korean Sign Language Augmentation(KoSLA) Corpus with Data Augmentation Technique,” in *Proceedings of the 2022 KIIT Autumn Conference*, Jeju, pp. 254-259, December 2022.

[10] J. W. Lee, “Autoethnography on Experience in Korean-Korean Sign Language Machine Translation Research,” *Korean Journal of Social Welfare*, Vol. 75, No. 4, pp. 145-172, November 2023. https://doi.org/10.20970/kasw.2023.75.4.005

[11] M.-S. Kim, M.-H. Yoo, J.-H. Jang, J.-Y. Choi, G.-E. Na, M.-S. Kim, and J.-J. Na, “Implementation of Deep Learning-Based Sign Language Acquisition Online Platform,” *Journal of Digital Contents Society*, Vol. 23, No. 11, pp. 2147-2157, November 2022. https://doi.org/10.

9728/dcs.2022.23.11.2147

[12] Enago Academy. Human Translation or Machine Translation, Which Is Better? [Internet]. Available: https://www.enago.co.kr/academy/machine-translators-vs-human-translators/.

[13] Kyung Hee University Graduate School News. Trends in AI Translation [Internet]. Available: https://khugnews.co.kr/?p=611/.

[14] J.-W. Kim and H.-Y. Jung, "End-to-End Speech Recognition Models Using Limited Training Data," *Phonetics and Speech Sciences*, Vol. 12, No. 4, pp. 63-71, December 2020. https://doi.org/10.13064/KSSS.2020.12.4.063

[15] Amazon Web Services. What are Large Language Models [Internet]. Available: https://aws.amazon.com/what-is/large-language-model/?nc1=h_ls/.

[16] H.-T. Joo, S. Lee, and K.-J. Kim, "Improving Korean Conversational Skills of LLaMA, a Public Large Language Model, Using Human and ChatGPT Conversations," in *Proceedings of the Korean Information Science Society Conference*, Jeju, pp. 991-993, June 2023.

[17] S. Shin, "Concept and Standardization Trends of Foundation Model of Supergiant AI," *Information and Communications Magazine*, Vol. 40, No. 6, pp. 12-21, May 2023.

[18] D.-J. Ji, H.-G. Jun, and D.-H. Im, "Spark-Based Big Data Preprocessing for Text Summarization," in *Proceedings of Annual Conference of KIPS 2022 (ACK 2022)*, Chuncheon, pp. 383-385, November 2022. https://doi.org/10.3745/PKIPS.y2022m11a.383

[19] S. Park, A Visual Analytics System for evaluating dataset of Neural Machine Translation, Master's Thesis, Seoul National University, Seoul, February 2023.

[20] J. Lim, "The Iconic Aspects and Semantic Properties of Korean Sign Language," *The Journal of Korean Language and Literature Education*, No. 68, pp. 63-88, October 2018. https://doi.org/10.17247/jklle.2018..68.63

[21] H. Kang and J. Yang, "The Analogy Test Set Suitable to Evaluate Word Embedding Models for Korean," *Journal of Digital Contents Society*, Vol. 19, No. 10, pp. 1999-2008, October 2018. https://doi.org/10.9728/dcs.2018.19.10.1999

[22] J. Lee, "A Purpose-Oriented Data Upstream and Downstream Strategy for Data-Intensive Society," in *Proceedings of 2019 Fall Conference of the Korea Technology Innovation Society*, Jeju, pp. 247-259, November 2019.

[23] M. Kim, J. Kim, and H. Y. Kim, "Sign2Gloss2Text-based Sign Language Translation with Enhanced Spatial-Temporal Information Centered on Sign Language Movement Keypoints," *Journal of Korea Multimedia Society*, Vol. 25, No. 10, pp. 1535-1545, October 2022. https://doi.org/10.9717/kmms.2022.25.10.1535

[24] H. C. Jeong, 2022 Korean-Korean Sign Language Parallel Corpus, National Institute of Korean Language, Seoul, NIKL 2023-01-19, April 2023.

[25] J. Seo, J. Kim, Y. Park, D. Kim, T. Kim, and M. Jeon, "A Study on Korean Sign Language Translation Script Method," in *Proceedings of the 23rd Convergence Conference of KASBA*, Seoul, pp. 191-198, August 2021.

주용민(Yong-Min Ju)

2019년 : 순천대학교 대학원
컴퓨터공학 (공학석사)
2023년 : 전남대학교 대학원
문화학 (박사수료)

2021년~현 재: 광주과학기술원 한국문화기술연구소 연구원
※관심분야 : 문화기술, 자연어처리, 알고리즘

김소진(So-Jin Kim)

2016년 : 조선대학교 대학원
디자인공학 (공학석사)
2022년 : 조선대학교 대학원
창의공학디자인융합학
(박사수료)

2018년~현 재: 광주과학기술원 한국문화기술연구소 연구원
※관심분야 : 서비스디자인, 사용자경험, 메타버스 등