한국 수어 글로스-문장 병렬 번역을 위한 전처리 전략과 성능 분석

강나연, 박영신, 장수현, 강지헌
덕성여자대학교
skdusb0007@duksung.ac.kr, p0tlsi@duksung.ac.kr, tngus7238@duksung.ac.kr, jhkang@duksung.ac.kr

A Study on Pre-processing Strategy and Performance Analysis for Korean Sign Language Gloss-Sentence Parallel Translation

Nayeon Kang, Youngshin Park, SuHyeon Jang, Jiheon Kang
School of Software, Duksung Women's University.

요 약

본 연구는 청각장애인의 정보 접근성 향상을 목표로 한국 수어 번역 파이프라인에서 글로스(gloss)를 한국어 문장으로 변환하는 단계에 집중한다. 수어와 한국어는 어순과 문법 체계가 상이하여 Sign2Text로 직접적인 번역이 어렵기 때문에 Sign2Gloss2Text 방식을 적용하여 수어 번역 정확도를 향상시켰다. 특히 번역 성능 향상을 위해 글로스 정규화, 형태소 기반 문장 토큰화, 데이터 정제 및 확장 등 다양한 전처리 기법을 적용하였다. 한국어 수어 번역 모델 구현을 위해 AI Hub의 재난 안전 정보 전달을 위한 수어영상 데이터 셋에서 글로스-문장 병렬 데이터를 기반으로 약 16만 건의 학습 데이터를 사용하여 트랜스포머 모델을 훈련하였다. 실험 결과 BLEU-4 점수 34.68을 기록하였으며, 이는 정제된 전처리 방식이 번역 품질에 크게 기여했음을 나타낸다. 본 연구는 수어 번역 시스템 개발의 초기 단계에서 전처리 전략의 중요성을 실증적으로 검증하였으며, 향후 수어 기반 번역 연구의 기초 자료로 활용될 수 있을 것으로 기대된다.

I. 서 론

청각장애인의 정보 접근성 개선을 위한 수어 자동 번역 기술에 대한 관심이 높아지고 있다. 수어는 고유한 문법과 어순을 가진 시각 언어로, 한국어와의 직접적인 기계 번역이 어렵다. 이에 따라 수어 번역 시스템은 일반적으로 파이프라인 구조를 취하는데, 이는 (1) 수어 영상 취득 → (2) 글로스(gloss) 맵핑→ (3) 텍스트(한국어 문장) 변환으로 이어지는 Sign2Gloss2Text 다단계 접근 방식이다.

본 연구는 이 중 글로스 → 텍스트 변환 구간에 초점을 맞추었다. 이 구간은 수어 특유의 단어 중심 구조에서 어미, 조사, 어순 등의 문법 요소가 중요한 한국어 문장으로 변환하는 핵심 단계다. 기존 수어 번역 연구들은 대부분 전체 파이프라인 또는 영상 입력 기반 번역에 초점을 두었지만, 본 연구에서는 데이터 정제와 언어 처리의 세밀함이 요구되는 글로스-문장 구간을 독립적으로 분석하였다.

특히, 본 연구는 모델 구조 자체보다는 학습 데이터 품질 향상과 전처리 전략이 번역 성능에 미치는 영향에 주목한다. 언어 번역에 자주 활용되는 트랜스포머 모델은 기존의 기계 번역 작업에서 이미 성능을 입증한 바 있으며, 본 연구에서도 이를 활용하되, 모델 설계보다는 데이터 구성과 정제에 집중하였다.

II. 본론

가. 트랜스포머 기반 언어 번역 모델

수화 번역을 위한 딥러닝 모델에는 주로 인공신경망 기반 기법들이 활용되어 왔다. 초기의 수화 번역 연구들은 순환신경망(RNN)과 어텐션 메커니즘을 조합하여 글로스 시퀀스를 음성 언어로 변환하거나 영상 입력을 직접 문장으로 변환하는 방법을 활용했다. 그러나 RNN 기반 모델은 긴 시퀀스 처리에서 한계가 있고 병렬화에 어려움이 있었다. Vaswani(2017) 등이 제안한 트랜스포머 모델은 어텐션 메커니즘을 기반으로 병렬 연산이 가능하면서도 장기간 의존관계를 효과적으로 학습할 수 있어 기계 번역 성능을 크게 향상시켰다. 트랜스포머는 이후 다양한 자연어 처리 과제에 표준 모델로 자리매김하며 수화 번역 분야에서도 이에 대한 활용이 보고되고 있다. 한승대 외(2022)는 트랜스포머를 이용한 한국 수어 글로스 변환 모델을 제안하여 기존 통계 기반 방법이나 RNN 기반 모델 대비 우수한 성능을 얻었음을 보고하였다.[1] 이처럼 트랜스포머 기반 접근은 한국어 수화 번역에 있어서도 효과적인 것으로 나타나고 있으며 본 연구에서는 이를 토대로 한국 수어 글로스-한국어 문장 번역 시스템을 구축하였다.

나. 데이터셋 및 전처리

AI Hub의 "재난안전 정보 전달을 위한 수어 영상 데이터셋"에서 제공하는 JSON 파일을 바탕으로 수어 글로스와 한국어 문장이 짝지어진 병렬 데이터를 기반으로 전처리/확장하여 별도 데이터세트를 구축하였으며. 원본 영상은 활용하지 않고, gloss_id 시퀀스와 한국어 문장 텍스트만을 사용하였다.

전처리는 다음과 같은 주요 과정을 거쳤다:

1. 데이터 필터링

해당 데이터세트는 자연재난, 사회재난, 날씨, 기타재난 분류에 의해 글로스-문장 쌍으로 구성되어 있으나, 일부 데이터 쌍의 경우 특정 시기(코로나19 유행기)의 제한적이고 반복적인 문장 구조가 포함되어 있어 일반 문장 구성과 어순의 차이로 인해 학습 편향을 유발하여 제외하였다.

2. 글로스 정규화

글로스는 '의미 기반 라벨링'이라는 특성상, 동일 의미의 수어와 동음이의어에 대한 수어를 구분하기 위해 단어 뒤에 숫자 접미어나 특수기호를 포함한 다양한 형태로 표기(예: '눈내리다1', '눈내리다2', '바람#')되어 있다. 이는 실제 수어 표현의 강도, 방향, 또는 기저 이미지에 따라 수어 영상을 구분하기 위함이나, 기계 번역 모델 관점에서는 동일 의미 단어가 서로 다른 토큰으로 인식되는 결과를 낳는다. 이로 인해 데이터 희소성과 학습 편향 문제가 발생하였다. 이에 본 연구에서는 숫자 접미사 및 특수기호를 제거하는 정규화 작업을 수행하여 의미 단위 기준의 일관된 토큰 구성

을 유도하였다. 김태용, 이명진, 김우영, 손의성, 김우주(2024)에 따르면 Space(공백)분절을 사용한 경우 BLEU-4 29.37을 기록하며 가장 뛰어난 성능[2]을 보인 반면, 본 연구에서는 글로스 정규화를 통해 BLEU-4 점수를 34.68까지 향상시켰으며, 이는 전처리 전략이 번역 품질에 미치는 영향이 매우 크다는 점을 실험적으로 입증하였다.

3. 글로스 토큰화

원본 JSON 파일에서 제공되는 gloss_id 리스트를 그대로 시퀀스로 활용하였다. AI-hub 데이터세트의 경우 글로스는 이미 의미 단위로 분절된 형태로 주어지므로, 별도의 토큰화 작업은 필요하지 않았다.

4. 한국어 문장 정제 및 토큰화

공백 기반 분절 방법은 음성언어를 공백 단위로 분할하는 전통적인 방법으로, 빠르고 간단하지만, 언어의 구조와 문맥을 고려하지 못할 수도 있다.[2] 이에 Mecab 형태소 분석기를 활용하여 문장을 의미 단위로 분해하였다.

5. 데이터셋 정리 및 확장

형태가 불분명하거나 중복된 의미를 가지는 문장을 제거하여 약 14만 건의 정제된 병렬 데이터를 확보하였다. 이후 모델의 일반화 성능 향상을 위해 직접 유사 문장을 수작업으로 생성하여 20,000건을 추가하고, 최종적으로 약 16만 개의 병렬 데이터를 확보하였다. 전체 데이터는 학습/검증/테스트 세트로 분할하여 과적합을 방지하였다.

특히 실제 데이터셋에 존재하지 않는 상황에 대응할 수 있도록, 의미는 유사하지만 표현이 달라질 수 있는 문장들을 중심으로 데이터 증강 작업을 수행하였다. 이는 단어 치환, 지명 변경, 반의어 활용, 품사 전환 등 다양한 방식으로 이루어졌으며, 모델의 문장 다양성 대응 능력을 높이는 데 기여하였다.

	글로스	한국어 문장
기존 데이터	오다 지역 비내리다 광주 일부분 오직	일부 광주 지역에 약한 비가 내리겠습니다.
추가 데이터	오다 지역 바람 대구 일부분 오직	일부 대구 지역에 약한 바람이 불겠습니다.
	오다 지역 서리 제주 일부분 오직	일부 제주 지역에 서리가 조금 내리겠습니다.
	오다 지역 눈내리다 속초 일부분 오직	일부 속초 지역에 눈이 조금 내리겠습니다.

〔표 1〕 데이터셋 증강 작업

III. 실험 결과

모델의 번역 성능은 표준 평가 척도인 BLEU(BiLingual Evaluation Understudy) 점수로 측정하였다.

		글로스	한국어 문장 번역 결과
예시 1	정답	전국에 구름이 많다가 점차 맑아지겠으나 동풍의 영향을 받는 강원도 영동 및 경북 북부 동해안은 대체로 흐리고 낮까지 비가 오겠습니다.	
	전처리 전	경북1 전국1 영동0 구름1 낮1 햇빛2 지도1 북쪽3 동쪽2 바다3 구름2# 비내리다1 바람1 까지1	전국 낮 구름 구름 2 2〈eos〉
	전처리 후	경북 전국 영동 구름 낮 햇빛 지도 북쪽 동쪽 바다 구름 비내리다 바람 까지	전국 이 대체로 맑겠으나 강원영동 과 경북 북부 동해안 은 동풍 의 영향 으로 점차 흐려져 낮 까지 비 가 오는 곳 이 있겠습니다
예시 2	정답	오늘 21시 강원, 경북 일부지역 한파, 야외활동 자제, 빙판길 낙상사고, 시설물관리 등에 유의바랍니다	
	전처리 전	신경쓰다1 밖1 부탁1 강원1 검사1 갑자기1 춥다1 축소1 사고1 온도내려가다1 경북1 조심1 미끄럽다1 행동1 일부분1 시:9시 얼음1 밤1 오늘1 길1 손목0 시설1	오늘〈unk〉강원, 경북 일부지역 한파,,, 야외 활동 자제, 빙판 길 낙상 사고, 시설물 관리 등 유의〈eos〉
	전처리 후	손목 일부분 조심 사고 갑자기 부탁 얼음 경북 온도내려가다축소 미끄럽다 밤 길 오늘 검사 강원 시:9시 밖 행동 시설 춥다 신경쓰다	오늘 밤 강원 (강릉) 지역 한파 , 낙상 사고 , 야외 활동 자제 , 빙판 길 낙상 사고 , 시설 물 관리 등 에 유의 바랍니다

〔표 2〕 전처리 전과 후 한국어 문장 번역 결과 비교

번역 예시를 보면, 제안 모델은 복잡한 수화 문장에 대해서도 한국어 어순에 맞게 단어를 배열하고 적절한 조사를 생성하는 등 자연스러운 문장을 출력하는 경향을 보였다. 다만, 문장 구조가 복잡하거나 문맥상 함축적인 표현의 경우 간혹 번역 오류나 부자연스러운 직역이 나타나기도 했다. 이는 주로 학습 데이터의 부족이나 수어 특유의 문법 요소(예: 높임법, 공간 활용)의 손실로 인한 것으로 판단된다.

IV. 결론

본 연구에서는 수어 자동 번역 단계 중 글로스(gloss)에서 한국어 문장으로의 변환 단계에 집중하여 정교한 전처리 전략이 번역 성능에 미치는 영향을 분석하였다. 수어와 한국어 간의 문법적 차이를 고려하여 트랜스포머 기반 신경망 모델을 적용하였으며, 특히 글로스 정규화, 형태소 기반 토큰화, 데이터 정제 및 수작업 확장과 같은 전처리 기법이 모델 성능 향상에 결정적인 역할을 함을 확인하였다. AI Hub의 수어 영상 데이터셋을 기반으로 약 16만 건의 병렬 데이터를 구축하고 모델을 학습한 결과, BLEU-4 점수 34.68이라는 우수한 번역 성능을 기록하였으며, 이는 수어 번역에서 데이터 품질과 정제 과정이 단순한 모델 구조 개선 못지않게 중요한 요소임을 보여주었다. 특히, 글로스 표기의 불일치 문제를 해결하기 위한 정규화 작업은 BLEU 점수를 두 배 이상 향상시키는 결과를 보이며 수어 번역 성능 개선에 있어 핵심적인 기여 요소로 작용하였다. 번역 예시를 통해 제안 모델이 자연스러운 어순 배열과 적절한 조사 생성을 수행하는 등 문장 생성 능력에서 높은 수준을 달성했음을 확인하였으며, 이는 실질적인 수어 번역 응용 가능성을 보여준다.

향후 영상 입력을 포함한 다중 입력 모델로의 확장을 통해, 보다 정교한 수어 번역 시스템 개발도 가능할 것으로 기대한다.

참 고 문 헌

[1] 한승대, 이현수, 김태환, 김주희. (2022), “트랜스포머를 이용한 한국 수어 글로스 변환 연구.” 한국정보과학회 2022 한국소프트웨어종합학술대회 논문집, 833 - 835.

[2] 김태용, 이명진, 김우영, 손의성, 김우주. (2024), ”수어 번역을 위한 글로스 분절방법 연구.“ 한국지능정보시스템학회 지능정보연구 제30권 제1호, 246-256.