JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 31 No. 4, pp. 33-44, April 2026
https://doi.org/10.9708/jksci.2026.31.04.033

Byte-Level Processing Limits in KSL Translation: A Study of the KoBART-ByT5 Performance Gap

Dong-Hyuk Kim*, Kyu-Cheol Cho**

*Student, Dept. of Computer Science & Engineering, Inha Technical College, Incheon, Korea
**Professor, Dept. of Computer Science & Engineering, Inha Technical College, Incheon, Korea

[Abstract]

This study investigates how input representation granularity affects performance and training behavior in Korean-to-Korean Sign Language(KSL) gloss translation. Using the National Institute of Korean Language Korean-KSL parallel corpus (2022-2024), we compare a token/subword-based pretrained Seq2Seq model (KoBART) with a byte-level model (ByT5). Quantitative results show a decisive advantage for KoBART, which achieves 0.447 METEOR versus 0.1192 (≈275% relative improvement[1]). Analyses indicate that ByT5 is constrained by substantially longer effective sequences, which degrades sentence-level generation, whereas KoBART benefits from subword segmentation that effectively performs structural alignment with KSL glosses, demonstrating superior suitability in terms of faithful information reconstruction. These findings provide empirical evidence for the critical role of input granularity design in low-resource KSL gloss translation and establish a robust baseline for future KSL machine translation studies.

▸ **Key words**: KSL Gloss Translation, Input Granularity, KoBART, ByT5, Seq2Seq Model, Transformer

[요 약]

본 연구는 한국어 문장을 한국수어(KSL) Gloss 시퀀스로의 번역에서, 입력 표현 단위가 성능과 학습 거동에 미치는 영향을 규명하고자 KoBART와 ByT5를 국립국어원 한국어-한국수어 병렬 말뭉치(2022-2024)로 비교하였다. 정량적 평가 결과, METEOR 지표에서 KoBART(0.447)가 ByT5(0.1192) 대비 약 2.7배 높은 성능을 기록하였다. 결과 분석에 따르면, ByT5는 입력 시퀀스 장기화로 문장 수준 생성에서 성능 한계를 보였다. 반면, KoBART는 Subword 기반 분절을 통해 수어 글로스와의 구조적 정렬을 효과적으로 수행함으로써 정보 재현 측면에서 우수한 적합성을 입증하였다. 본 연구는 저자원 환경인 KSL Gloss 번역에서 입력 단위 설계의 중요성을 실증하고, 향후 수어 기계 번역 연구를 위한 정량적·정성적 기초 자료를 제공한다는 점에서 학술적 의의가 있다.

▸ **주제어**: 한국수어 글로스 번역, 입력 표현 단위, 토큰 기반 사전학습 모델, 바이트 단위 사전학습 언어 모델, 시퀀스 투 시퀀스 모델, 트랜스포머

1) Relative Improvement (%) = (0.447 - 0.1192) / 0.1192 × 100 ≈ 275.0%

• First Author: Dong-Hyuk Kim, Corresponding Author: Kyu-Cheol Cho
*Dong-Hyuk Kim (dh95229505@gmail.com), Dept. of Computer Science & Engineering, Inha Technical College
**Kyu-Cheol Cho (kccho@inhatc.ac.kr), Dept. of Computer Science & Engineering, Inha Technical College
• Received: 2026. 01. 26, Revised: 2026. 02. 24, Accepted: 2026. 03. 31.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

한국수어(이하 KSL, Korean Sign Language)는 한국어를 손동작으로 단순 치환한 보조수단이 아니라, 한국어와 구별되는 고유한 문법 체계와 어휘 체계를 갖는 독립된 언어로 이해되어야 한다[1]. 이와 같은 언어적 특성으로 인해, 공공, 안전, 교육, 의료 등 정보가 집중되는 영역에서 수어 기반 정보 접근성은 기술적, 제도적 과제로 대두된다. 특히 국가 승인 통계에 기반한 2023년 '한국수어 활용 조사'[2]에 따르면, 장애 정도가 심한 청각장애인 중 30.1%가 수어를 주된 의사소통 방법으로 사용하며, 수어 통역 서비스가 가장 필요하다고 응답한 영역은 의료기관(83.0%)이 가장 높은 응답률을 보였다. 또한 병원에서 원하는 의사소통 지원으로 '수어 가능한 직원 배치'(92.5%)가 제시되어, 의료 환경에서 수어 접근성에 대한 요구가 구체적인 형태로 제기되고 있음을 시사한다. KSL은 한국어와는 상이한 문법 구조를 가질 뿐만 아니라, 전형적인 저자원(low-resource) 언어의 특성을 보인다. 이러한 데이터 희소성은 기계 번역 모델이 언어 간의 구조적 대응 관계를 학습하는 데 큰 장애 요인으로 작용한다.

본 연구는 최신 수요와 자원 확장 흐름을 배경으로, 한국어->KSL gloss(수어 의미 라벨) 생성 과정에서 입력 표현 단위(input granularity) 의 설계가 모델의 성능과 학습 거동에 미치는 영향을 심층적으로 분석한다. 이를 위해 subword 기반 시퀀스 투 시퀀스(Sequence-to-Sequence, Seq2Seq) 모델인 KoBART(Korean Bidirectional and Auto-Regressive Transformers)와 byte-level 모델인 ByT5(Byte-level Text-to-Text Transfer Transformer)를 대상으로 실험을 수행하며, Optuna를 통한 하이퍼파라미터 최적화를 적용하여 각 모델의 잠재 성능을 공정하게 평가한다.

본 연구의 주요 기여는 다음과 같다.

- 한국어-KSL gloss 번역 과업에서 입력 단위(Subword vs Byte-level)에 따른 성능 차이를 실증적으로 규명하고, 데이터 희소성 환경에서의 모델 선택 기준을 제시함.
- BPE(Byte Pair Encoding) 기반 KoBART와 byte-level(ByT5) 접근법 간의 학습 수렴 패턴 차이를 분석하고, 출력 시퀀스 길이 증가가 의미 정렬 능력에 미치는 영향을 규명함.
- 기존 서술 중심의 분석을 넘어, 하이퍼파라미터 최적화와 정량·정성 분석을 결합하여 저자원 수어 번역 연구를 위한 체계적인 검증 프레임워크를 제공함.

본 논문의 구성은 다음과 같다. II장에서는 배경 개념과 관련 논의를 정리하고, III장에서는 비교 설계와 최적화 절차를 제시한다. IV장에서는 정량적 분석과 정성적 분석을 통해 두 접근의 성능 및 학습 거동 차이를 해석하며, V장에서 결론과 향후 연구 방향을 논의한다.

II. Preliminaries

2.1 Existing Approaches

KSL gloss 번역의 여러 선행 연구는 토큰 기반 번역 모델을 채택해 왔다[3][4]. 이러한 흐름은 한국어-수어 병렬 코퍼스가 제한적이라는 데이터 조건에서, Subword 단위를 활용한 토큰 기반 접근은 입력을 비교적 안정적으로 분절할 수 있다는 실질적 이점을 제공해 왔다[5][4]. 특히 BPE 기반 어휘집 구성은 고빈도, 중빈도 어휘를 압축적으로 표현함으로써 학습 계산을 효율화하고[6], 정규화된 subword 단위는 학습 안정성을 보조하여 한국어와 gloss 간 의미 단위 정렬의 일관성을 유지하는 데 기여할 수 있다[3][5].

그러나 토큰 기반 구조는 몇 가지 구조적, 언어학적 한계를 내포한다. subword 분절은 단어 수준 OOV(Out-of-Vocabulary)를 일부 완화하지만, KSL-gloss의 표제부 기호나 특정수형(handshape) 표기 같은 특수 표기는 토크나이저 학습 분포에서 벗어나기 쉽다. 이때 특수 표기는 과도한 분절을 초래하고, 결과적으로 표기 coverage gap이 남는 제약으로 이어진다[6]. 또한 희귀 단어나 표기 변이에 취약하여 예측 오류를 범할 수 있다는 점이 지적되어 왔다[6]. 이와 같은 한계는 한국어-수어 gloss의 복잡한 형태적 특성, 말뭉치 희소성, 그리고 표기 체계 비표준성에 의해 더욱 심화되며, 결과적으로 토큰 기반 접근만으로는 저자원 환경에서 의미 표현의 견고성을 충분히 확보하기 어렵다는 점을 시사한다.

최신 연구 동향에서는 이러한 저자원 수어 번역 한계를 극복하기 위해, 대형 언어 모델(Large Language Model, LLM)에 기호에 대한 자연어 설명(description)을 매핑하는 동적 프롬프팅 기법(AulSign) 등이 2025년에 제안된 바 있다[19]. 그러나 거대 모델과 문맥 학습(in-context learning)에 의존하는 거시적 접근 방법과는 독립적으로, 기계 번역 모델이 원시 코퍼스를 어떠한 해상도 단위(Granularity)로 처리하는지가 정보 정렬 및 디코딩 성능에 미치는 본질적 영향에 대한 실증 연구는 여전히 요구된다.

2.2 Byte-level Models for KSL Gloss

바이트 단위로 입력을 처리하는 Byte-level 모델은 최근 여러 자연어 처리 연구에서 특히 저자원 환경에서 강력한 성능을 보이는 접근 방식으로 주목받고 있다[6]. Byte-level 접근은 subword/word-level 어휘집 의존을 최소화하고, 입력을 바이트 시퀀스로 직접 모델링함으로써 token-free 입력 표현을 구성한다. 그 결과, 토큰 기반 모델의 어휘집 기반 토크나이징에서 불가피하게 발생하는 <unk> 및 어휘 확장 부담을 구조적으로 회피할 수 있으며, 희귀 어휘, 특수 기호, 형태 변형이 빈번한 비정규 텍스트도 안정적으로 인코딩할 수 있다는 장점이 보고되고 있다[6].

이러한 byte-level 입력 처리의 어휘집 비의존성, 비정형 표기 견고성, 그리고 저자원 학습 효율성은 KSL이 지닌 언어적 조건과도 밀접하게 맞닿아 있다. KSL gloss 말뭉치는 규모가 제한적이며, 정제되지 않은 표기 변이가 빈번히 관찰되는 저자원 데이터의 성격을 가진다. 또한 gloss 표기 방식은 대문자 표기, 약어형 gloss ID, 부호형 기호 등이 혼재하는 경우가 많아, 토큰 기반 분절 과정에서 표기 불일치가 확대될 수 있다. 이때 byte-level 입력은 표기 변이를 문자적 수준에서 직접 포착함으로써, 토크나이저의 어휘집 경계가 유발하는 coverage gap을 상대적으로 완화하는 방향으로 작동할 수 있다.

단, byte-level 입력은 동일 문장을 subword 단위보다 더 미세한 단위로 전개하므로, 입력 시퀀스 길이를 증가시키는 경향이 있다.

2.3 KoBART vs. ByT5: Input Representation Granularity Trade-off

본 연구는 동일한 Transformer encoder-decoder 계열 내에서 입력 표현 단위 차이가 생성 품질과 학습 거동에 미치는 영향을 분리하여 논의한다. 이 관점에서 비교하는 두 모델의 핵심 대비점은, subword 기반 분절이 학습, 추론 효율성을 제공하는 반면 byte-level 입력은 표기 변이 및 희귀 기호에 대한 강건성을 제공하되[6], 바이트 시퀀스의 길이 증가로 인해 학습, 추론 비용이 상승할 수 있다는 trade-off로 요약될 수 있다[5]. 그러나 한국어-KSL gloss 번역 맥락에서, 이러한 trade-off 중 어느 요인이 실제 생성 품질과 학습 거동을 지배하는지는 선험적으로 결정하기 어렵다. 따라서 본 연구는 byte-level(ByT5)과 subword-level(KoBART) 중 어떤 입력 단위가 KSL gloss 번역에서 더 우수한 생성 품질을 제공하며, 그 차이가 학습 거동에서 어떻게 설명되는가를 단일 실증 질문으로 설정한다. 이를 위해 본 연구는 KoBART와 ByT5를 비교 대상으로 설정하고, 두 입력 표상의 구조적 차이가 KSL gloss 생성 성능과 학습 특성에 미치는 영향을 실험적으로 검증한다.

III. Experimental Setup for KSL Gloss Translation Validation

본 장에서는 한국어-KSL gloss 번역 작업에서 입력 표현 단위 차이가 Seq2Seq 모델의 학습 및 생성 특성에 어떤 영향을 미치는지 규명하기 위해, 토큰 기반 KoBART와 바이트 기반 ByT5를 동일한 조건하에서 비교하는 실험 설계를 제시한다.

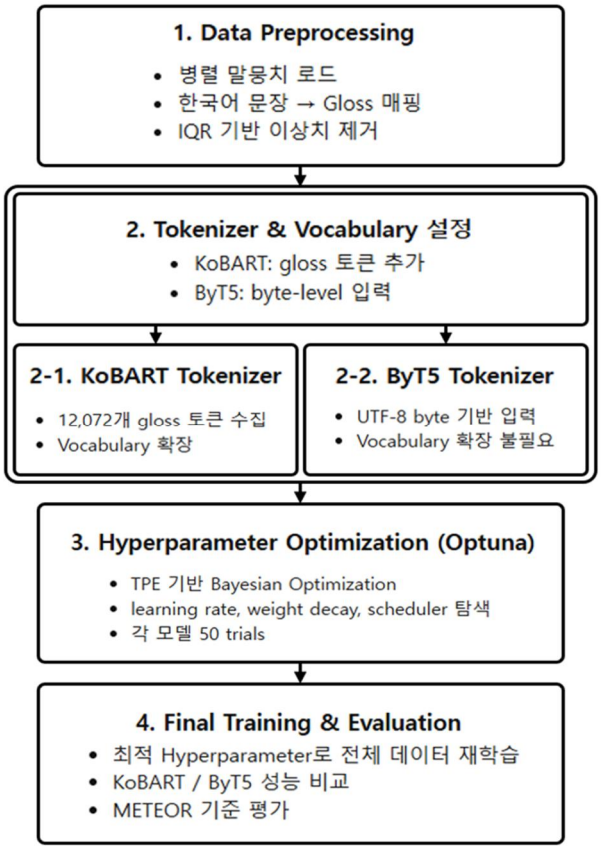

Fig. 1. Experiment Design

제안하는 설계([Fig.1])는 병렬 말뭉치 구성 및 데이터 정제, 입력 표현 단위가 실제로 달라지는 지점인 토크나이저 커스텀, 어휘 처리 단계, Optuna 기반 하이퍼파라미터 최적화, 최적 설정으로의 재학습과 최종 평가로 구성된다. 동시에, 자동 평가 지표를 분석 기준으로 별도 정리하여 다음 장의 성능 분석이 동일한 설계 위에서 일관되게 해석될 수 있도록 기반을 마련한다. 이에 따라 이후 절에서는 데이터

및 정제 기준, 모델별 입력 표현과 어휘 구성, 평가 지표 정의, 그리고 학습 최적화 기법을 순서대로 기술한다.

3.1 Data Source and Preprocessing Strategy

3.1.1 Data Source

본 연구에서는 국립국어원에서 제공하는 공개 수어 병렬 말뭉치(Korean Sign Language Parallel Corpus, 2022-2024)를 사용하였다[7]. 각 샘플은 한국어 문장과 해당 수어 영상의 gloss 정보로 구성되며, gloss 정보는 시작 시각, 끝 시각, gloss_id 등으로 이루어진다([Fig 2]).


```
{
"id" : "VXPAKOKS240768550",
  ...
  "krlgg_sntenc" : {
    "koreanText" : "예비 신부는 예물 선물로 ...",
          },
"sign_script" : {
    "sign_gestures_strong" : [ {
      "start" : 2.098,
      "end" : 2.355,
      "gloss_id" : "아내1",
      ...
    }, ...
}
```


Fig. 2. Raw Data Structure

그러나 학습에는 모든 메타데이터가 필요하지 않으므로, 본 연구에서는 학습에 직접적으로 활용되는 핵심 요소인 koreanText와 gloss_id만을 추출하였다. 해당 추출을 통해 모델이 문장-수어 gloss 간의 직접적인 매핑 관계를 학습하도록 하였다.

3.1.2 Data cleaning

전처리된 데이터는 한국어 문장과 이에 대응하는 gloss 리스트 쌍으로 구성되어 있다. 그러나 원시 데이터 분석 결과, 일부 문장에서 한국어 문장 길이와 gloss 리스트 길이의 비정상적인 불균형이 확인되었다. 이는 데이터 수집 과정에서의 심각한 정렬 오류나 무의미한 부연 설명이 과도하게 포함된 노이즈로 분석된다. 이러한 입력 및 출력 길이 불균형은 학습 시 모델이 구조적 의미보다는, 길이 편향에 과적합할 위험을 증가시킨다[8]. 특히 byte-level 모델은 시퀀스 길이 증가 폭이 상대적으로 커질 수 있다는 점을 고려할 때 그 영향이 상대적으로 확대될 수 있다[5]. 불균형을 정량적으로 확인하기 위해, 본 연구에서는 공백 기준 단어 수를 채택하였다. 이는 교착어인 한국어 특성상 형태소 단위 분석에 비해 정교함은 다소 부족할 수 있으나, 두 언어 간의 상대적인 길이 불균형을 파악하고 이상치를 판별하는 목적 달성에는 해당 기준이 충분히 타당한 척도로 작용한다.

한국어 문장의 단어 수 (len(korean_text.split())와 gloss 리스트의 gloss 수를(len(gloss_ids))를 각각 계산하여, 두 길이의 비율(Length Ratio = len(gloss)/len(koreanText))을 산출하였다. 산출된 자료에서 이상치 판단을 위해 Tukey의 사분위 범위(Interquartile Range, IQR) 규칙을 적용하였다. IQR 규칙은 탐색적 데이터 분석에서 널리 사용되는 통계적 방법으로, 데이터의 제1사분위수(Q1)와 제3사분위수(Q3)를 이용해 IQR을 (1)식과 같이 계산한다[9][10].

$$IQR = Q3 - Q1 \quad \cdots\cdots(1)$$

$Q1$: 제1사분위수 $Q3$: 제3사분위수

이후, 상·하한 수염 (Upper Whisker, Lower Whisker)은 식 (2)에 따라 정의되며, 이를 초과하거나 미만인 값을 이상치로 간주한다.

$$\begin{aligned} Upper\ Whisker &= Q3 + 1.5 \times IQR \\ Lower\ Whisker &= Q1 - 1.5 \times IQR \end{aligned} \quad \cdots\cdots(2)$$

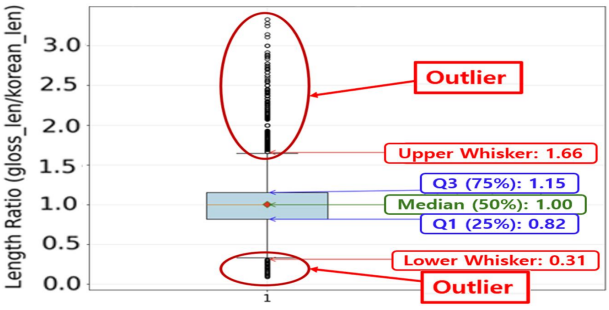

Fig. 3. Boxplot of Length Ratio

[Fig.3]의 Box plot으로 시각화한 결과 상한 수염은 1.66으로 계산되었으며, 하한 수염은 0.31로 계산되었다. 이는 gloss 리스트의 길이가 한국어 문장보다 1.66배 이상 길거나 0.31배 이하로 짧을 경우를 이상치로 간주할 수 있음을 의미한다.

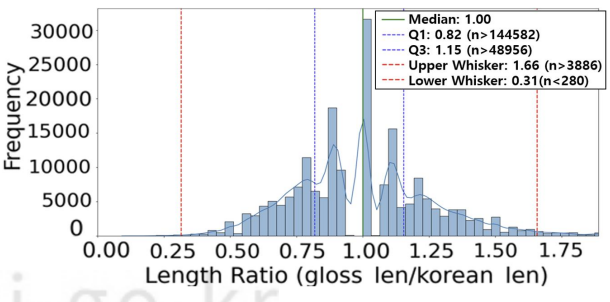

Fig. 4. Histogram of Length Ratio

상기 기준으로 전체 데이터의 분포를 히스토그램으로 나타낸 결과([Fig. 4]), 데이터의 중앙값은 1.00이었으며 제1사분위수(Q1)는 0.82, 제3사분위수(Q3)는 1.15로 대다수의 관측치가 해당 정상 범주에 안정적으로 밀집해 있음을 확인하였다. 데이터 품질을 향상 시키고, 학습 과정에서 노이즈가 모델 성능에 영향을 미치는 것을 최소화하기 위해 IQR 기반 상·하한 수염을 벗어나는 샘플 총4,166건을 학습에서 선제적으로 제외하였다.

분석 결과([Fig. 4]), 제외된 샘플은 하한 수염 미달인 280건과 상한 수염을 초과하는 3,886건으로 세분된다. 이상치 제거 과정을 거쳐 최종적으로 본 실험에 활용된 데이터셋의 구체적인 규모와 한국어 문장 및 수어 글로스(Gloss)의 통계적 특성은 [Table 1]과 같다. 추가로, 제거된 4,166건의 샘플이 특정 도메인에 편향되어 있는지 확인하기 위해 7개 주요 도메인(일상생활, 금융, 교육, 방송, 의료, 관광, 민원 행정)에 대한 분포를 분석하였다. 분석 결과, 제거된 샘플의 도메인 분포는 전체 데이터셋의 분포와 최대 0.46%p 이내의 근소한 차이만을 보였으며, 이를 통해 본 정제 과정이 특정 주제나 표기 유형에 편향되지 않은 객관적 절차임을 검증하였다.

Table 1. Detailed Dataset Statistics and Sequence Properties

Category	Count	Mean		Var	
		Kor	Gloss	Kor	Gloss
Raw Corpus	168,651	11.31	10.18	4.82	13.92
Final Dataset	164,485	11.29	10.32	4.74	9.14
Train Split	131,587	11.29	10.32	4.74	9.14
Validation Split	16,449	11.29	10.32	4.74	9.14
Test Split	16,449	11.29	10.30	4.73	9.15

특히, [Table 1]에서 제시하듯 훈련 및 검증에 사용되는 데이터(148,036건)의 규모는 본 번역 과제에 존재하는 전체 고유 수어 글로스(Gloss ID)의 총 개수(12,072개)와 대비해 볼 때, 개별 글로스당 평균적으로 약 12.3문장의 예문만이 주어지는 극단적인 데이터 희소성을 보인다. 이러한 수치는 KSL이 지닌 저자원 환경의 한계를 정량적으로 증명하며, 본 연구가 탐구하는 '제한적인 조건에서 입력 단위에 따른 일반화 성능'에 대한 실험적 타당성을 부여한다.

3.2 Tokenization Strategy and Vocabulary Customization

KoBART는 subword 기반 토크나이저를 사용하기 때문에 수어 gloss 토큰을 직접 생성하기 위해서는 학습 데이터에 해당하는 gloss ID가 어휘집에 포함해야 한다. 평가 데이터의 정보가 학습 과정에 사전 노출되는 정보 누출을 원천적으로 차단하기 위해, 오직 학습 데이터에서 관찰된 고유 gloss ID(12,072개)만 추출하여 KoBART의 어휘집을 확장하였다. 이러한 확장은 모델이 수어 gloss의 고유 표기 체계를 인식하고 문맥 내에서 적절히 디코딩할 수 있도록 돕는다.

ByT5의 경우 Byte-level 토크나이저를 사용하기 때문에, 별도의 어휘 확장이나 커스터마이징 과정이 필요하지 않다[6]. 결과적으로, 두 모델의 토크나이징 전략은 gloss 입력을 표현하는 방식에서 근본적인 차이를 만들어내며, 해당 차이는 이후 실험에서 두 모델의 성능과 수렴 패턴 차이를 발생시키는 중요한 요인으로 작용한다.

3.3 Evaluation Metrics

본 연구는 생성된 KSL gloss의 품질을 정량적으로 평가하기 위해, 기계 번역 분야에서 널리 활용되는 BLEU(Bilingual Evaluation Understudy), ROUGE(Recall-Oriented Understudy for Gisting Evaluation), METEOR(Metric for Evaluation of Translation with Explicit ORdering) 세 가지 자동 평가 지표를 사용하였다. 각 지표는 토큰 단위의 정밀도(BLEU)와 시퀀스의 구조적 보존(ROUGE-Lsum), 그리고 의미론적 정렬(METEOR)이라는 다각적인 관점에서 상호 보완적으로 작용한다. 본 연구에서는 특히 수어 글로스의 기호적 특성과 어순 유동성을 효과적으로 포착할 수 있도록 이와 같은 다차원적 평가 체계를 구성하였으며, 이 중 의미 정렬에 강점이 있는 METEOR를 주 성능 지표로 선정하였다.

3.3.1 BLEU

BLEU는 기계 번역 텍스트의 품질을 평가하기 위한 대표적인 자동 평가 지표로, 기계 번역 결과가 인간 번역(reference translation)과 얼마나 유사한지를 정량적으로 측정한다. 기계 번역 평가는 생성 결과가 인간 번역문과 유사할수록 번역 품질이 우수하다는 전제에 기반한다. BLEU 지수는 번역의 수정된 n-그램 정밀도(n-gram precision, p_n)의 기하 평균과 번역 길이가 적절한지 확인하는 간결성 페널티(Brevity Penalty, BP)를 결합하여 0에서 1 사이의 값(3)이 산출된다[11].

$$BLEU = BP * \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \cdots\cdots\cdots\cdots (3)$$

- BP : 간결성 페널티(Brevity Penalty)
- w_n : 각 n-gram의 중요도를 나타내는 가중치 (Weight)
- p_n : 수정된 n-gram 정밀도 (Modified n-gram Precision)

후보 번역이 참조 번역에 대해 얼마나 많은 n-그램을 공유하는지를 평가한다. 따라서 단어 선택 및 어순이 유사할수록 높은 점수를 얻는다.

그러나 BLEU는 문자열 유사성에만 초점을 맞추어 의미적 유사성을 충분히 반영하지 못하고, 동의어나 대체 표현 사용 시 점수가 왜곡될 수 있다는 한계가 있다. 또한, BLEU는 참조 번역의 수와 품질에 민감하고, 과적합 된 시스템의 경우 실제 번역 품질과 무관하게 높은 점수가 부여될 가능성이 있다[12]. 이러한 특성 때문에 BLEU는 본 연구에서 보조 지표로 활용되며, 주로 단기적 패턴 정합 여부를 확인하기 위한 용도로 사용한다.

3.3.2 ROUGE

ROUGE는 모델이 기준 텍스트의 주요 내용과 얼마나 유사한지, 그리고 중요한 정보를 얼마나 잘 포착했는지를 평가한다. ROUGE-N은 후보 출력과 참조 출력들 사이의 n-gram recall로 정의되며[13], 참조 출력에 등장한 n-gram을 후보 출력이 얼마나 포착했는지를 측정한다.

$$ROUGE-N=\frac{\sum_{r\in R}\sum_{gram_n\in r} Count_{match}(gram_n)}{\sum_{r\in R}\sum_{gram_n\in r} Count(gram_n)} \cdots\cdots(4)$$

- $\sum_{r\in R}\sum_{gram_n\in r} Count_{match}(gram_n)$: 후보 번역에서 참조 번역과 정확히 일치한 n-gram의 총수
- $\sum_{r\in R}\sum_{gram_n\in r} Count(gram_n)$: 참조문장 n-gram 전체 총량
- R: 하나의 입력에 대한 참조 번역의 집합
- r: 집합 R의 원소로, 개별 참조 번역 1개 (단일 문장)
- $gram_n$: r에서 추출되는 연속된 n개 단위로 구성된 n-gram

한편 ROUGE-Lsum은 최장 공통부분 수열(Longest Common Subsequence: LCS)에 기반하여 두 시퀀스의 유사도를 계산한다. ROUGE-Lsum은 두 문장 간의 LCS를 이용하여 단어의 연속적인 일치뿐만 아니라 비연속적인 일치도 평가하므로, 보다 유연하게 의미적 유사성을 반영할 수 있다[13]. 결과적으로, BLEU는 참조 문장과 단어, 구문 수준의 표현이 얼마나 유사한지를 중심으로 평가하는 반면, ROUGE는 생성 문장이 참조 문장의 핵심 내용이 얼마나 포함되었는지를 평가하여, 두 지표는 평가 관점에서 상호 보완적으로 활용될 수 있다. 수어 gloss의 경우 문장이 매우 압축적이고 핵심 의미만 나열되는 경향이 있으므로, ROUGE-Lsum은 gloss 번역 작업에서 정보 보존의 정도를 측정하는 데 적합한 보조 지표로 작용한다.

3.3.3 METEOR

METEOR는 BLEU가 가진 일부 문제를 해결하고, 문장 또는 세그먼트 수준에서 인간의 판단과 더 높은 상관관계를 보이도록 설계된 기계 번역 평가 지표이다. 평가는 문장을 단어 수준으로 분해한 1-gram 단위의 요소인 unigram 단위를 사용한다. 문장 내 단어(unigram)의 일치 정도를 precision과 recall의 조화 평균인 F_{mean}을 식 (5)와 같이 산출하며, 통상적으로 recall에 더 높은 가중치 (α)를 부여한다[14].

$$F_{mean}=\frac{P\times R}{\alpha P+(1-\alpha)R} \cdots\cdots(5)$$

$$P=\frac{\text{일치한 } unigram \text{ 수}}{\text{후보문장 내 } unigram \text{ 수}},\ R=\frac{\text{일치한 } unigram \text{ 수}}{\text{참조문장 내 } unigram \text{ 수}}$$

그러나 F_{mean}은 단어의 순서를 고려하지 않으므로, METEOR는 정렬된 단어들이 얼마나 연속적으로 배치되었는지를 나타내는 chunk의 수를 통해 이를 보완한다. 청크가 많을수록 높은 Penalty가 부과되며, 이 벌점을 F_{mean}에 반영하여 식 (6)과 같이 최종 점수를 산출한다[14].

$$METEOR=(1-Penalty)\times F_{mean} \cdots\cdots(6)$$

$$Penalty=0.5\times\left(\frac{chunk \text{ 수}}{\text{일치한 } unigram \text{ 수}}\right)^3$$

METEOR는 단순한 정확 단어 일치뿐만 아니라 어근 일치, 동의어 일치, 어순 페널티 등을 포함하여 단어의 형태적 변형이나 의미적 유사성까지 반영할 수 있다.

결과적으로 METEOR는 BLEU보다 의미적 유사성을 더 정확히 포착하며, 단어 형태 변화나 어순 차이가 있어도 보다 인간 판단에 가까운 평가를 제공한다[14].

KSL-Gloss는 표기자에 따라 다양한 동의어가 사용될 수 있으며, 어순이 한국어와 다르며 유동적이라는 특성을 지닌다[15]. 이러한 특성을 고려하면, 동의어 매칭과 단편화 페널티를 통해 어순 변이를 부분적으로 허용하면서 의미 정렬을 평가할 수 있는 METEOR[14]가 gloss 생성 품질 평가에 적합하다. 따라서 본 연구에서는 METEOR를 주 평가 지표로 선정하였으며, BLEU, ROUGE는 gloss의 표면 수준 및 정보 단위의 보완적 평가를 위해 보조 지표로 사용하였다.

3.4 Controlled Training Protocol

두 모델 간 성능 차이가 순수하게 아키텍처 특성에서 비롯되도록 모든 조건을 동일하게 통제하였다. 정제된 샘플을 train 80%, validation 10%, test 10%로 분할하였으며, 재현성 보장을 위해 고정 시드값을 적용하였다.

본 실험에서 고려된 각 모델의 파라미터 규모, 입출력 사양, 배치 크기 및 메모리(VRAM) 점유량 등의 세부 명세는 [Table 2]와 같다.

Table 2. Computational Budget and Model Configurations

Metric	KoBART	ByT5
Model Parameters	~124 M	~582 M
Max Input/ Target Length	256	256
Batch Size	128	128
VRAM Usage	~16 GB	~44 GB

두 모델의 비교에서 하이퍼파라미터 설정은 성능을 좌우하는 주요 요인이므로 동일한 탐색 예산과 동일한 최적화 절차를 적용하여 공정한 조건에서 최적 설정을 도출하고자 한다. 이를 위해 hyperparameter 최적화에는 Bayesian Optimization 기법을 기반으로 하는 Optuna 라이브러리를 사용하였다[16]. 탐색 알고리즘으로는 TPE(Tree-structured Parzen Estimator) 기반으로 탐색을 수행하여 이전 trial의 성능을 바탕으로 탐색 공간을 점진적으로 개선하는 구조를 가진다.

IV. Experimental Results and Analysis

본 장에서는 동일한 학습 환경과 전처리 조건에서 수행된 실험을 기반으로, KoBART와 ByT5가 한국어-KSL gloss 번역 작업에서 보이는 성능적 차이를 정량적으로 분석한다. 이를 통해 KSL 번역 작업에서 입력 표현 방식이 성능 결정에 미치는 근본적 영향뿐 아니라, 모델 선택 시 고려해야 할 실질적 기준을 제시한다.

4.1 Hyperparameter Optimization Outcome

본 연구는 Optuna 라이브러리를 통해 KoBART, ByT5의 구조적 차이에 따라 민감도가 다르게 나타나는 hyperparameter 조합을 효율적으로 도출하였다. 탐색 시간 단축을 위해 전체 학습 데이터 및 검증 데이터의 20%만을 사용하여 총 50회의 trial을 수행하였다. 각 trial은 검증 손실 및 meteor 점수를 기준으로 평가하였으며, 그 결과 도출된 최적의 하이퍼파라미터는 [Table.3]과 같다.

Table 3. Optimal parameters by mode

Hyperparameter	KoBART	ByT5
learning rate	4.486×10^{-5}	4.538×10^{-5}
weight decay	0.0484	0.0645
generation num beams	2	2
lr scheduler type	constant	polynomial
optimizer	adafactor	adamw_torch
warmup ratio	0.0720	0.0085

4.2 KoBART Model Performance

[Fig.5]는 epoch에 따른 METEOR 변화를 시각화한 결과로, Validation METEOR가 학습 초반(1~5 epoch)에 급격히 상승한 뒤 10~15 epoch 전후에 약 0.45 수준으로 도달하며 안정적으로 수렴하는 경향을 나타낸다.

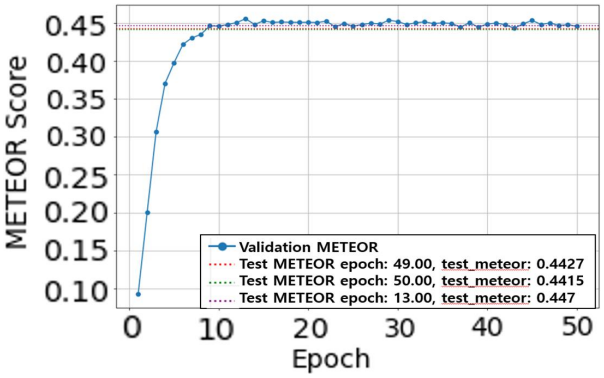

Fig. 5. METEOR Scores Across Epochs for the KoBART Model

한편 Test METEOR는 epoch 13에서 0.447로 최고점을 기록한 이후 epoch 49는 0.4427, 50은 0.4415 수준으로 소폭 하락하여, 검증 지표의 안정적 수렴이 유지되더라도 일반화 성능 관점에서 epoch 13 부근이 최적 체크포인트임을 시사한다. 이러한 METEOR의 안정적 수렴은 KoBART가 의미적 유사성 평가에서 견고한 일반화 능력을 갖춘다는 점에서 중요한 결과이다.

참조 문장과의 단어, 구문 중복 정도를 반영하는 BLEU 지표를 기준으로 [Fig. 6]을 분석한 결과, Validation 값은 학습 초반(1~10 epoch)에 급격히 상승한 뒤 약 0.20 내외에서 빠르게 수렴하며, epoch 증가에 따라 추가적인 향상은 관찰되지 않았다. Test 값 역시 epoch 13에서 0.2008, epoch 49에서 0.2042로 차이가 0.0034에 불과하여, 이후 추가 학습이 단어, 구 수준의 표면 정합을 개선 폭이 매우 제한적임을 시사한다.

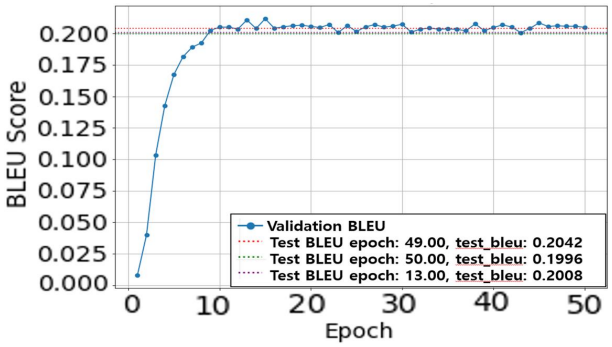

Fig. 6. BLEU Scores Across Epochs for the KoBART Model

내용 보존과 핵심 정보 유지 정도를 평가하는 ROUGE-Lsum 지표([Fig. 7])를 기준으로 분석한 결과, Validation 값은 학습 초반(1-5 epoch)에 약 0.75 수준에서 0.78대 후반까지 급격히 상승하였다. 이후 50 epoch까지 0.79 내외의 좁은 범위에서 소폭 변동하며 점진적으로 포화되는 경향이 관찰되었다. 이는 KoBART가 비교적 짧은 학습 구간에서 이미 문장 수준의 핵심 중복 구조를 학습하고, 이후 추가 학습이 정보 보존 관점에서 개선 폭이 매우 제한적임을 의미한다.

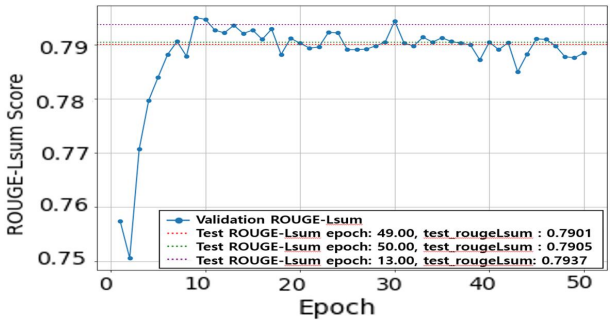

Fig. 7. ROUGE-Lsum Scores Across Epochs for the KoBART Model

BLEU 및 ROUGE-Lsum 은 10-15 epoch 이후 유의미한 변동 없이 수렴하며, METEOR 기반 분석과 동일한 추세가 확인되었다. 지표 전반에서 일관된 수렴 양상이 관찰되어 모델 학습의 안정성이 뒷받침된다.

KoBART 모델의 일반화 성능은 Loss 분석에서 더욱 명확하게 드러난다. Training Loss는 전체 epoch 동안 꾸준히 감소하였지만, Validation Loss는 약 10~15 epoch에서 최저점(≈0.27)을 기록한 후 완만히 증가하였다. 이는 10~15 epoch 이후 모델이 과적합 되기 시작했음을 나타낸다.

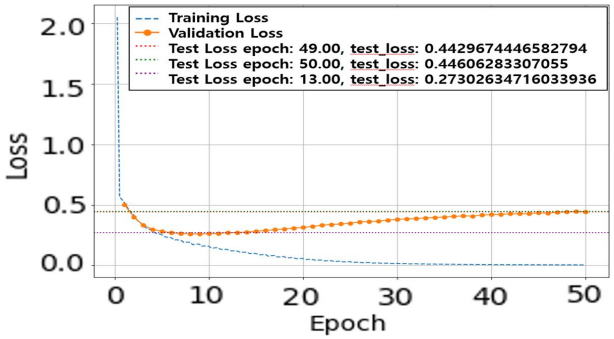

Fig. 8. Loss Values Across Epochs for the KoBART Model

Validation Loss, Test Loss, METEOR 의 변화를 함께 고려할 때, epoch 13에서 최적 성능이 관찰되었다. 이는 KoBART가 중간 단계의 학습에서 가장 강한 일반화 성능을 발휘하며, 장기 학습이 오히려 성능 저하를 유발할 수 있음을 의미한다.

4.3 ByT5 Model Analysis of Structural Limits

ByT5 모델의 METEOR 성능은 학습 초반에 빠르게 증가하는 패턴을 보였으나, 전체 성능의 절댓값은 KoBART 대비 현저히 낮게 유지되었다. Validation METEOR는 1~5 epoch 구간에서 약 0.05 수준에서 출발하여 급격한 상승을 보였고, 10 epoch 전후에 0.11~0.115 수준까지 도달하였다. 이후 학습이 진행되면서 20 epoch 부근에서 0.118~0.122 범위로 수렴하며 미세한 진동을 반복하는 형태를 나타냈다. 특히 30 epoch 이후로는 지표 변동 폭이 극히 좁아지며 모델이 학습할 수 있는 패턴을 대부분 포착한 것으로 해석된다. 그러나 지표의 절댓값이 매우 낮다는 점은, ByT5가 byte-level 입력을 기반으로 하여 의미 단위보다 작은 하위 단위에 집중함으로써 문장 수준 의미 정렬 능력이 제한됨을 의미한다[5][17].

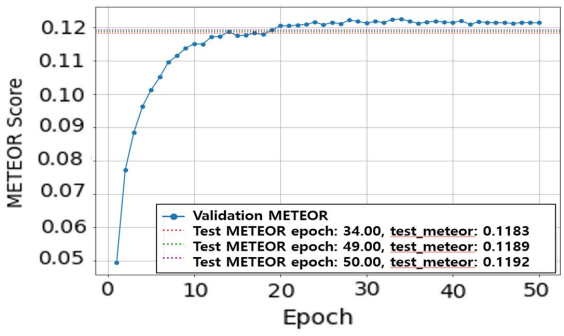

Fig. 9. METEOR Scores Across Epochs for the ByT5 Model

ByT5의 BLEU 점수는 전체 지표 중 가장 낮은 절댓값을 보였다. Validation BLEU는 초기 학습 단계에서 약

0.001 수준에서 시작하여 10 epoch 부근에서 0.013까지 증가하였으나, 이후 0.013~0.014 범위에 머무르며 정체 상태에 진입하였다. 이는 ByT5가 표면적 n-gram 정렬 기반 지표에서는 안정성을 확보하고 있으나, byte 단위의 입력 특성 때문에 어휘 단위 의미 구조를 포착하는 능력이 매우 제한적임을 반영한다.

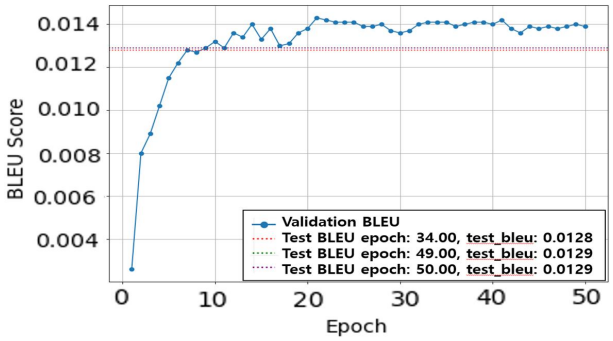

Fig. 10. BLEU Scores Across Epochs for the ByT5 Model

ROUGE-Lsum은 학습 전체에서 0.327~0.333 범위 내에서만 미세하게 진동하는 양상을 보였다. 그래프상의 변동은 불규칙하게 나타날 수 있으나, 변동 폭이 약 ±0.003 범위에 머물러 수렴 안정성이 높은 것으로 해석된다. ROUGE-Lsum이 LCS 기반 재현율을 측정한다는 특성을 고려할 때, 해당 미세 진동은 byte-level 인코딩이 문장 내 연속적 의미 구조를 안정적으로 반영하기 어려운 한계에서 비롯된 것으로 판단된다[5][18].

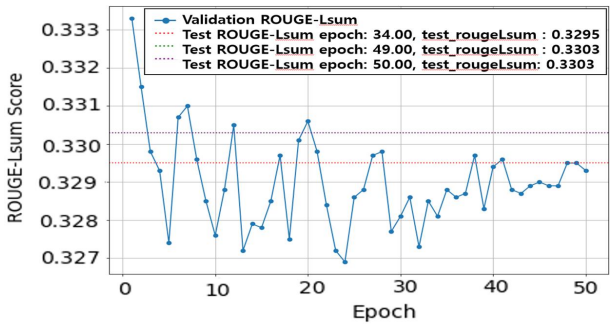

Fig. 11. ROUGE Scores Across Epochs for the ByT5 Model

Test ROUGE-Lsum 역시 각 체크포인트가 동일한 분포를 보이며, ByT5가 정보 재현 능력 자체는 일정 수준 유지하되 절대 성능은 매우 낮음을 의미한다.

ByT5의 가장 뚜렷한 학습 특성은 Loss의 조기 포화(Early Saturation) 현상이다([Fig.12]). Training 및 Validation Loss 모두 1~2 epoch 내에 매우 급격히 감소하여 0에 가깝게 수렴하였고, 이후 거의 변하지 않는 패턴이 확인되었다. 이는 한국어 문자가 UTF-8 인코딩 시 여러 바이트로 분해되기에, 모델이 의미 단위에 대비해 미세한 byte-level 반복 패턴을 상대적으로 빠르게 학습할 수 있기 때문이다[6][17].

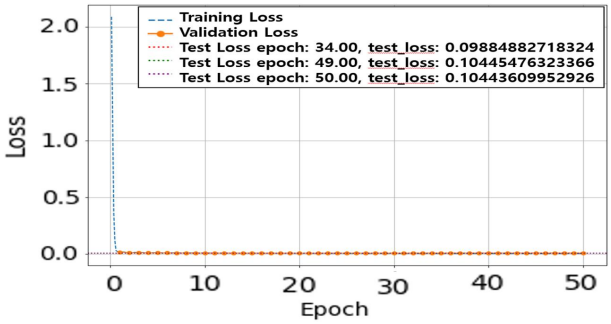

Fig. 12. Loss Values Across Epochs for the ByT5 Model

그러나 Loss는 매우 낮은 값으로 포화되었음에도 불구하고, METEOR, BLEU, ROUGE-Lsum 모두 극히 낮은 절대 성능을 기록하였다. [Fig. 12]에서 확인된 추세는 byte-level 모델이 문자 수준 오류 감소에는 강점을 보이지만, gloss 번역에 필요한 문장 수준 의미 정렬 능력 학습에는 구조적 한계가 있는 것으로 판단된다[5]. 즉, ByT5는 학습 안정성과 입력 강건성에서는 강점을 가지지만, 의미 기반 번역에는 적합하지 않은 구조적 한계를 내재하고 있다.

4.4 Comparative Performance Analysis

KoBART와 ByT5를 비교한 결과, KoBART는 METEOR, BLEU, ROUGE-Lsum 등 주요 자동 평가 지표 전반에서 ByT5보다 일관되게 더 높은 성능을 보였다[Table 4]. 특히 gloss 번역의 의미 정렬을 대변하는 핵심 지표로 METEOR를 기준으로 평가할 경우, KoBART(0.447)와 ByT5(≈0.12) 간 격차가 두드러지며, 다른 지표에서도 동일한 방향성의 성능 차이가 관찰된다. 반면 ByT5는 학습 전 구간에서 지표 변동 폭이 작고 Loss가 1~2 epoch 내 조기 포화되는 등 수렴 안정성은 높지만, 손실이 낮음에도 번역 지표가 낮게 유지되는 loss-metric 불일치가 지속되어 문장 수준 의미 정렬을 충분히 학습하지 못한 것으로 해석된다[5][6]. 또한, 두 모델은 각각 13 epoch(KoBART)와 50 epoch(ByT5)라는 비교적 짧은 학습 구간 내에서 빠르게 수렴하는 경향을 보였다. 이러한 조기 수렴 현상은 대규모 Pre-trained Language Model의 강력한 가중치가 특정 좁은 도메인(KSL 번역) 과업에 빠르게 적응하는 전형적인 fine-tuning 패턴으로 분석된다.

Table 4. Quantitative Comparison of KoBART and ByT5 on Korean-KSL Gloss Translation

Metric	KoBART (Epoch 13)	ByT5 (Epoch 50)	KoBART - ByT5
METEOR	0.447	0.1192	+0.3278
BLEU	0.2008	0.0129	+0.1879
ROUGE-Lsum	0.7937	0.3303	+0.4634
Test Loss	0.243	0.1044	-0.1386

정량적 수치와 함께, 이러한 성능 격차가 통계적으로 유의한지 검증하기 위해 Bootstrap Resampling (n=100,000)을 수행하였다. 검증 결과, 리샘플링 과정에서 ByT5의 평균 성능이 KoBART를 상회하는 경우가 단 한 차례도 관찰되지 않아 p-value < 0.00001 수준에서 매우 유의한 성능 차이가 있음이 확인되었다. 이는 두 모델의 성능 차이가 단순한 표본 오차나 우연에 의한 것이 아니라, 입력 표현 단위의 구조적 차이에 따른 근본적인 성능 격차임을 시사한다.

이와 같은 결과는 동일한 조건에서 단순히 어휘 확장 여부에 따라 나타난 현상이라기보다, ‘Token-based’와 ‘Token-free’라는 상이한 아키텍처 설계 철학의 차이에서 비롯된 것으로 해석할 수 있다. KoBART의 subword(BPE) 분절은 의미 단위에 보다 근접한 중간 표현을 제공함으로써, gloss와 같이 의미 단위 간 대응이 중요한 생성 과제에서 정렬 부담을 완화하고 정보 재현의 안정성을 높인다[14][17]. 반면 ByT5는 UTF-8 기반의 byte-level 입력을 사용하므로 유효 시퀀스 길이가 크게 증가하고[6], 이에 따라 self-attention이 처리해야 하는 상호작용이 급증하여 주의 집중이 분산되며 의미 단위 수준의 추상화가 어려워진다[5][17]. 또한 서브워드 기반 모델이 어휘집을 통해 유효한 의미 단위를 명시적으로 구조화하는 반면, ByT5는 바이트 단위로부터 직접 의미를 학습하도록 설계되어 어휘집에 의한 구조적 제약을 두지 않는다. 결과적으로 본 비교는 저자원 KSL 환경에서 subword 기반 KoBART가 byte-level ByT5보다 의미 정렬 및 정보 재현 측면에서 구조적으로 더 적합함을 시사한다.

4.5 Qualitative Error Analysis

정량적 평가 지표가 포착하지 못하는 두 모델 간의 실제 번역 품질 차이와 오류 발생 원인을 파악하기 위해 정성적 오류 분석(Error Typology)을 수행하였다. 테스트 셋에서 추출한 98건의 샘플을 다음의 3가지 주요 오류 유형으로 기준을 세워 분류하였다.

1. 의미론적 유사성(Semantic): 못된 단어나 문맥상 유사한 의미의 다른 단어를 선택하는 오류
2. 구문론적 정렬(Alignment): 수어 Gloss 고유의 배열 및 어순이 어긋나는 오류
3. 정보 누락/중복(Content): 핵심 단어가 과도하게 생략되거나 무의미하게 반복되는 오류이다.

Table 5. Distribution of Error Types

Error Type	KoBART	ByT5
Semantic	93	82
Alignment	3	13
Content	2	3

분석 결과, KoBART의 오류는 대다수 문장의 뼈대와 문법 구조를 유지한 채 일부 단어가 유사한 의미의 다른 단어로 대체되는 의미론적 오류에 집중되어 있었다. 반면 ByT5는 KoBART 대비 구문론적 정렬 오류가 눈에 띄게 높게 나타났다. 두 모델의 질적 격차가 극명하게 드러나는 대표 사례는 [Table 6]와 같다.

Table 6. Distribution of Error Types

Source	Sentence
Korean Text	여의도는 주말엔 괜찮은데 , 평일에는 특히 이 시간대엔 교통이 막힐 수 있어
Reference	여의도2, 곳1, 매주1, 결국1, 괜찮다1, 보통1, 차막히다1, 때2, 지시1#
KoBART	여의도2, 곳1, 매주1, 결국1, 괜찮다1, 보통1, 차막히다1, 때2, 지시1#
ByT5	여의도2, 토요일1, 일요일2, 괜찮다1, 보통1, 때2, 특별1, 지시1#, 때2, 교통1, 막히다2, 가능1

이를 종합할 때, KoBART의 높은 정량적 수치가 실제 생성 품질의 구조적 강건성으로 직결됨을 시각적으로 알 수 있다.

V. Conclusions

본 연구는 한국어-KSL gloss 번역 작업에서 입력 표현 방식이 서로 다른 KoBART(subword-based)와 ByT5(byte-level)를 동일한 데이터 전처리 및 최적화 방식 아래에서 비교하였으며, KoBART가 더 높은 과제 적합성을 가진다는 결론을 도출하였다. 특히 KoBART는 의미 정렬과 구조적 대응 측면에서 더 안정적으로 gloss를 산출하였다. 자동 평가 지표 전반에서 KoBART가 일관된 우위를 보였으며, 학습 과정에서도 출력 품질과 직접 연결되는 의

미 단위 학습이 더 견고하게 형성되는 양상이 관찰되었다. 반면 byte-level 접근은 의미 정렬과 구조 대응 측면에서 상대적으로 불리한 경향을 보였다. 본 비교 실험의 결과는 KSL gloss 번역 성능이 입력 표현 단위가 제공하는 의미 단위 정렬 가능성에 의해 좌우되며, 순서, 구조 대응이 중요한 gloss 생성에서는 subword 분절이 실질적 이점을 제공할 수 있음을 시사한다. 그러나, 본 연구는 표준 모델 구조를 기준으로 비교 분석을 수행하였으므로, ByT5의 시퀀스 길이 증가로 인한 성능 저조를 완화할 수 있는 Sparse Attention, 입력 길이 제한과 같은 구조적 개선 기법을 직접 통합하여 평가하지 못했다는 한계가 있다. 이러한 요인 완화 설정과의 비교 실험은 향후 연구에서 비정형 입력의 잠재력을 추가 검증하기 위한 후속 과제로 남긴다. 또한 향후 연구에서는 subword와 byte fallback을 결합한 hybrid tokenization 전략을 도입하여, 의미 정렬 능력과 OOV 처리 강건성을 동시에 확보할 수 있는 방향도 함께 검토할 필요가 있다.

REFERENCES

[1] National Institute of Korean Language, “Sign Language Introduction,” National Institute of Korean Language, *https://www.korean.go.kr/front/page/pageView.do?mn_id=202&page_id=P000300.*

[2] Ministry of Culture, Sports and Tourism (MCST), National Institute of Korean Language (NIKL), “Announcement of Results of the ‘2023 Korean Sign Language Usage Survey’ (Press Release),” National Institute of Korean Language, *https://www.korean.go.kr/front/board/boardStandardView.do?b_seq=958&board_id=6.*

[3] W. Kim, T. Kim, B. Kim, M. Lee, G. Lee, K. Kim, J. Cha, and W. Kim, “SSL: Korean Disaster Safety Information Sign Language Translation Benchmark Dataset,” Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pp. 9948-9953, Torino, Italy, May 2024. DOI: N/A.

[4] T.-V. Dang, G.-H. Yu, J.-Y. Kim, Y.-H. Park, C.-W. Lee, and J.-Y. Kim, “Korean Text to Gloss: Self-Supervised Learning approach,” Smart Media Journal, Vol. 12, No. 1, pp. 32-46, Feb. 2023. DOI: 10.30693/SMJ.2023.12.1.32.

[5] J. Libovický, H. Schmid, and A. Fraser, “Why don’t people use character-level machine translation?,” Findings of the Association for Computational Linguistics: ACL 2022, pp. 2470-2485, Dublin, Ireland, May 2022. DOI: 10.18653/v1/2022.findings-acl.194.

[6] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel, “ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models,” Transactions of the Association for Computational Linguistics, Vol. 10, pp. 291-306, Mar. 2022. DOI: 10.1162/tacl_a_00461.

[7] National Institute of Korean Language (NIKL), “Modu Corpus: Korean-Korean Sign Language Parallel Corpus (2022, 2023, 2024) - Request Page,” Language Information and Resources (Modu Corpus), *https://kli.korean.go.kr/corpus/main/requestMain.do.*

[8] D. Variš and O. Bojar, “Sequence Length is a Domain: Length-based Overfitting in Transformer Models,” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 8246-8257, Online and Punta Cana, Dominican Republic, Nov. 2021. DOI: 10.18653/v1/2021.emnlp-main.650.

[9] The Pennsylvania State University (Penn State), Eberly College of Science, “3.2 - Identifying Outliers: IQR Method,” STAT 200: Elementary Statistics, *https://online.stat.psu.edu/stat200/lesson/3/3.2.*

[10] J. W. Tukey, “Exploratory Data Analysis,” Addison-Wesley Publishing Company, pp. 39-44, 1977. ISBN: 9780201076165.

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, Philadelphia, Pennsylvania, USA, July 2002. DOI: 10.3115/1073083.1073135.

[12] C. Callison-Burch, M. Osborne, and P. Koehn, “Re-evaluating the Role of Bleu in Machine Translation Research,” Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL), pp. 249-256, Trento, Italy, Apr. 2006. DOI: N/A.

[13] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” Proceedings of the Text Summarization Branches Out, pp. 74-81, Barcelona, Spain, Jul. 2004. DOI: N/A.

[14] S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65-72, Ann Arbor, Michigan, USA, Jun. 2005. DOI: N/A.

[15] Y.-M. Ju and S.-J. Kim, “Study of the Production of Korean Sign Language Scripts and Preprocessing of Text Data for Machine Translation Targeting Deaf Individuals,” Journal of Digital Contents Society, Vol. 25, No. 10, pp. 2829-2840, Oct. 2024. DOI: 10.9728/dcs.2024.25.10.2829.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2019), pp. 2623-2631, Anchorage, Alaska, USA, Aug. 2019. DOI: 10.1145/3292500.3330701.

[17] U. Shaham and O. Levy, “Neural Machine Translation without Embeddings,” Proceedings of the 2021 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021), pp. 181-186, Online, Jun. 2021. DOI: 10.18653/v1/2021.naacl-main.17.

[18] S. Yavuz, C.-C. Chiu, P. Nguyen, and Y. Wu, "CALCS: Continuously Approximating Longest Common Subsequence for Sequence Level Optimization," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), pp. 3708-3718, Brussels, Belgium, Oct.-Nov. 2018. DOI: 10.18653/v1/D18-1406.

[19] L. Bulla, G. Tuccio, M. Mongiovì, and A. Gangemi, "Leveraging Large Language Models for Accurate Sign Language Translation in Low-Resource Scenarios," Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025), pp. 4217-4224, Bologna, Italy, Oct. 2025. DOI: 10.3233/FAIA251315

Authors

Dong-Hyuk Kim received the Associate of Science (A.S.) and the Bachelor of Science (B.S.) degree in Computer Science & Engineering from Inha Technical College, Korea, in 2025 and 2026 respectively. His research interests include medical artificial intelligence, multi modal, and Natural Language Processing.

Kyu-Cheol Cho received the B.S., M.S. and Ph.D. degrees in Computer Science and Information Engineering from Inha University, Korea, in 2005, 2007 and 2013, respectively. Dr. Cho joined the faculty of the Department of Computer Science & Engineering at Inha Technical College, Incheon, Korea, in 2016. He is currently a assistant professor in the Department of Computer Science & Engineering, Inha Technical College. He is interested in cloud computing, green IT and web programming.