

다국어 COIL에서 전공용어 일관성 확보를 위한 RAG 기반 번역 시스템 설계

Design of a RAG-Based Translation System for Terminology Consistency in Multilingual COIL

최만규¹, 이선희^{2*}

Minkyu Choi¹, Sunhee Lee^{2*}

요약

다국가 대학 간 Collaborative Online International Learning(COIL)에서는 참여자 다수가 영어 비 원어민이며 서로 다른 모국어를 사용하는 경우가 많다. 이 환경에서 일반 기계번역(Machine Translation, MT)이나 범용 대규모 언어 모델(Large Language Model, LLM)을 사용하면 동일 전공용어가 문장마다 다르게 번역되어 협업 소통에 장애가 생길 수 있다. 본 연구는 강의자료·용어집을 벡터 데이터베이스에 색인하고, 번역 시 관련 맥락을 검색하여 LLM에 제공하는 Retrieval-Augmented Generation(RAG) 기반 번역 시스템을 설계한다. 기술적으로는 기존 용어 제약 기계번역(terminology-constrained MT) 및 용어집 주입 연구의 연장선에 있다. 본 연구의 기여는 다자·비영어권 COIL이라는 교육 맥락에 특화된 점과, 번역 품질과 함께 ‘전공용어 일관성 비율’을 평가 지표로 도입한 점이다. 프로토타입은 Python·FastAPI로 구현하며, 평가 설계에서는 일반 MT, LLM 단독, 용어집 전체 정적 삽입, RAG(검색 기반)의 네 조건을 비교하도록 구성하였다.

핵심어 : COIL, 전공용어 일관성, RAG, 용어 제약 기계번역, 다국어 협업학습, 시스템 설계

Abstract

In multilingual Collaborative Online International Learning (COIL) settings, many participants are non-native English speakers who also speak different first languages. In such settings, general Machine Translation (MT) or general-purpose Large Language Models (LLMs) often render the same specialized term inconsistently across sentences, which can hinder collaborative communication. This paper designs a Retrieval-Augmented Generation (RAG)-based translation system that indexes course materials and glossaries in a vector database and retrieves relevant context at translation time to guide an LLM. Technically, the system builds on prior work on terminology-constrained machine translation and glossary injection. Its contributions are its specialized application to multi-party, non-English-mediated COIL contexts and its introduction of a terminology consistency ratio alongside translation quality as an evaluation metric. A prototype is implemented in Python and FastAPI, and the evaluation design compares four conditions:

1 SW Convergence Education Center, Mokwon University, Daejeon, Korea [Senior Researcher]

e-mail: mk.choi@mokwon.ac.kr

2 Department of Computer Engineering, Mokwon University, Daejeon, Korea [Assistant Professor]

e-mail: sunnylee@mokwon.ac.kr

Received(August 2, 2026), Review Result(1st: August 19, 2026, 2nd: September 9, 2026), Accepted(October 13, 2026), Published(October 31, 2026)



© 2026 The Authors. Published by NCISS.

This is an open access article licensed under the Creative Commons Attribution-NonCommercial 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

commercial MT, LLM-only, static full-glossary insertion, and the proposed RAG approach.

Keyword : COIL, terminology consistency, RAG, terminology-constrained MT, multilingual collaborative learning, system design

1. 서론

고등교육의 국제화는 물리적 이동을 전제로 한 전통적 교류에서, 디지털 기술을 활용한 가상 국제 협업으로 확대되고 있다. Collaborative Online International Learning(COIL)은 서로 다른 국가의 교육기관이 온라인으로 공동 학습 활동을 설계·운영하는 교수학습 모형으로, 학생들이 해외 이동 없이도 국제적 협업 경험과 문화 간 역량을 함양할 수 있게 한다 [1][2].

COIL이 다국가 네트워크로 확장되면서 참여 학생의 언어 배경도 다양해지고 있다. 상당수가 영어 비원어민이며, 영어를 공통 매개어로 쓰기 어려운 상황이나 비영어권 언어 간 직접 소통이 필요한 경우가 발생한다. 가상 교환·COIL 연구에서 언어 능력의 불균형과 의사소통 장벽은 반복적으로 보고되는 도전 과제이다. 특히 전공 수업에서 사용하는 전문용어가 기계번역이나 범용 LLM에 의해 문장마다 다르게 번역되면, 학습자 간 의미 공유가 어려워질 수 있다. 다만 ‘용어 불일치가 COIL 협업 실패로 이어진다’는 인과를 직접 검증한 실증 연구는 아직 제한적이며, 본 연구는 그 간극을 메우기 위한 시스템 설계와 평가 프레임워크를 제시하는 데 목적이 있다.

기존 AI 활용 COIL 연구는 주로 양자 간 단기 프로젝트에서 생성형 AI를 문화적 촉진자나 일반 학습 보조 도구로 활용하는 데 초점을 맞추어 왔다. 영어를 공통 매개어로 전제하는 경우가 많으며, 전공용어의 일관된 번역을 시스템적으로 보장하는 설계는 드물다. 본 연구는 교과 맥락(강의자료·용어집)을 검색하여 LLM 번역에 반영하는 RAG 기반 시스템을 설계하되, 기술 자체보다는 ‘다자·비영어권 COIL’이라는 응용 맥락에 특화해 적용하는 데 초점을 둔다.

본 연구의 목적은 다음과 같다. 첫째, 다국어 COIL 환경에서 전공용어 일관성을 지원하기 위한 시스템 요구사항을 도출한다. 둘째, 기존 용어 제약·용어집 주입 연구의 연장선에서 교과 맥락 검색을 결합한 번역 파이프라인을 설계한다. 셋째, 일반 MT, LLM 단독, 용어집 정적 삽입, RAG의 네 조건을 비교하는 평가 설계를 제시한다.

연구 질문은 다음과 같다.

- RQ1. 다국어 COIL 환경에서 전공용어의 일관된 번역을 지원하기 위한 시스템 요구사항은 무엇인가?
- RQ2. 교과 지식베이스 검색을 결합한 번역 파이프라인은 어떻게 설계될 수 있으며, 전체 용어집 정적 삽입과 어떤 차이가 있는가?
- RQ3. 네 가지 조건(일반 MT/LLM 단독/용어집 정적 삽입/RAG)에서 번역 품질과 전공용어 일관성은 어떻게 비교·평가될 수 있는가?

2. 이론적 배경 및 관련 연구

2.1 COIL과 확산 동향

COIL은 2000년대 중반 이후 본격적으로 개념화되었으며, 서로 다른 국가의 교수자가 공동으로 학습 활동을 설계하고 학생들이 온라인으로 협업하는 것을 핵심으로 한다 [1]. 최근 스코핑 리뷰에 따르면 2010~2024년 실증 연구만 90편 이상이 확인되며, Parallel, Consultation, Jigsaw, Hybrid 등 다양한 구현 모델이 존재한다 [2]. COIL은 문화 간 역량·디지털 리터러시·글로벌 협업 기술 함양에 긍정적 영향을 미치는 것으로 보고되나, 시차·제도적 지원·평가 방식 등의 과제도 지속적으로 지적된다 [2][3].

다자 네트워크로 확장될수록 언어적·문화적 이질성이 커진다. 가상 교환·COIL 문헌에서는 언어 능력의 불균형, 의사소통 장벽이 참여와 학습에 영향을 미친다는 점이 여러 연구에서 지적된다. 본 연구는 그중에서도 ‘전공용어의 일관된 공유’에 초점을 맞춘다.

2.2 AI를 활용한 COIL 연구 동향

생성형 AI의 확산과 함께 COIL에 AI를 접목하려는 시도가 증가하고 있다. 기존 연구들은 주로 AI를 토론 촉진, 아이디어 생성, 문화적 관점 제시, 언어 지원 등의 보조 도구로 활용한다. 일부 연구에서는 AI 지원 COIL이 문화 간 역량과 언어 자신감 향상에 기여할 수 있음을 시사하나, 대부분 양자 간·단기·소규모이며 영어를 공통 매개어로 전제하는 경우가 많다.

본 연구는 AI를 단순 보조 도구가 아니라 ‘교과 맥락을 반영한 번역 지원 엔진’으로 위치시킨다. 기술적 신규성보다는, 다자·비영어권 COIL이라는 응용 맥락에 특화된 점과 전공용어 일관성을 명시적 설계·평가 목표로 삼은 점에 기여가 있다.

2.3 신경망 기계번역과 전문용어 처리

신경망 기계번역(Neural Machine Translation, NMT)은 높은 유창성을 달성하였으나 도메인 불일치와 전문용어 처리에서 취약점을 보인다 [4]. 전문용어는 특정 분야에서 합의된 대응 관계가 있어야 하며, 동일 용어가 문장마다 다르게 번역되면 전문 소통의 신뢰성이 저하된다.

용어를 제어하기 위한 연구는 이미 활발하다. Post와 Vilar는 어휘 제약을 강제하는 디코딩 알고리즘을 제안하였고 [5], Dinu 등은 용어 제약을 학습 단계에서 주입하는 방법을 제시하여 추론 속도 저하 없이 용어 준수율을 높였다 [6]. WMT Terminology Shared Task는 용어 사전을 활용한 번역 성능을 체계적으로 평가하는 벤치마크를 제공하였다 [7]. 저자원·비영어 언어쌍에 대해서는

NLLB(No Language Left Behind) 계열의 대규모 다국어 모델이 기반을 제공하고 있다 [8].

상용 MT와 범용 LLM은 일반 문장에서는 우수하나, 특정 교과의 용어집·강의자료에 정의된 용례를 반영하지 못한다. COIL처럼 반복적 협업 소통이 이루어지는 환경에서는 이 한계가 의미 공유를 저해할 수 있다. 본 연구는 위 선행 연구의 연장선에서, ‘교과 단위 지식베이스 검색’을 결합하는 방식을 COIL 맥락에 적용한다.

2.4 RAG와 용어집 기반 번역

Retrieval-Augmented Generation(RAG)은 사전 학습된 생성 모델과 외부 지식 저장소를 결합하는 기법이다 [9]. 기계번역·용어 제어 맥락에서도 용어집을 검색해 프롬프트에 삽입하는 접근이 연구되고 있다. 다만 RAG 자체를 새로운 기술로 주장하기보다는, 이미 확립된 용어 제약 기계번역과 용어집 주입 연구의 한 변형으로 위치시키는 것이 타당하다.

본 연구에서는 일반 웹 문서가 아니라 해당 교과의 강의자료·용어집·이전 협업 산출물을 지식베이스로 구축하고, 번역 요청 시 관련 구문·용어 정의를 검색하여 LLM에 제공한다. 핵심 질문은, 용어집 전체를 프롬프트에 정적 삽입하는 방식과 검색으로 필요한 부분만 제공하는 방식 중 무엇이 더 효과적인가이다. 용어집과 강의자료가 커질수록 정적 삽입은 컨텍스트 길이 제한과 노이즈 증가에 직면하므로, 검색의 필요성이 커진다.

3. 시스템 설계

3.1 설계 요구사항 도출

[표 1] 시스템 설계 요구사항

[Table 1] System Design Requirements

구분	요구사항 내용
기능적	강의자료·용어집 업로드 및 색인
	번역 요청 시 관련 맥락(용어 정의·용례·문단) 자동 검색
	검색 맥락을 반영한 용어 일관성 유지 번역 생성
	다국어 쌍 지원 및 용어집 다국어 대응 관계 반영
	번역 결과와 함께 사용된 용어 대응 정보(출처) 제시
비기능적	준실시간 응답(검색 지연과 정적 삽입 대비 트레이드오프 관리)
	지식베이스 갱신 용이성
	접근 통제 및 최소 보관 원칙을 통한 개인정보·저작권 보호

본 연구의 시스템 설계 요구사항은 Peffers 등의 설계과학연구방법론(Design Science Research Methodology, DSRM)에서 제시한 문제 식별 - 해결목표 정의 - 설계 및 개발의 논리를 참고하였다 [10]. 다국어 COIL 환경의 전공용어 불일치, 교과 맥락 반영의 어려움, 다국어 용어 대응 및 검색 효율성을 주요 문제로 설정하고, 이를 해결하기 위한 기능적·비기능적 요구사항으로 구성하였으며, 그 내용은 [표 1]과 같다.

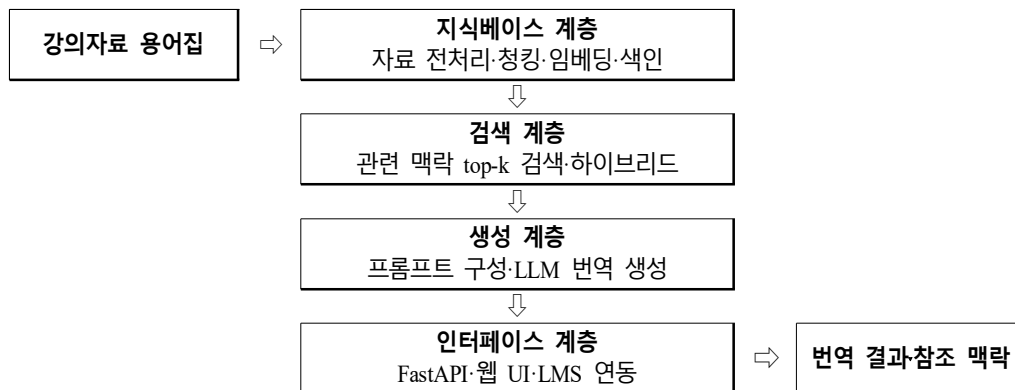
3.2 전체 아키텍처

시스템은 [표 2]와 같이 네 계층으로 구성된다. 지식베이스 계층에서 강의자료와 용어집을 색인하고, 검색 계층에서 관련 맥락을 선별한 뒤, 생성 계층에서 LLM을 통해 번역을 수행하며, 인터페이스 계층이 사용자와 연동한다. 전체 구조는 [그림 1]에 제시하였다.

[표 2] 시스템 아키텍처 계층 구성

[Table 2] System Architecture Layers

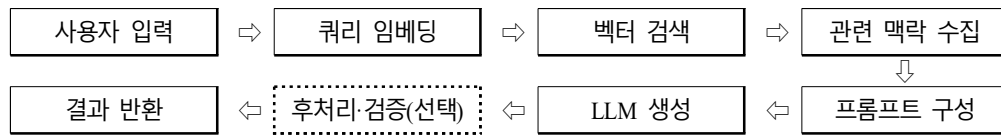
계층	주요 기능
지식베이스 계층	강의자료·용어집 전처리, 청킹, 임베딩, 벡터DB 색인 및 메타데이터 관리
검색 계층	쿼리 임베딩, 상위 k개 관련 청크 검색, 키워드·의미 하이브리드 검색
생성 계층	검색 맥락+원문+대상 언어+일관성 지시로 프롬프트 구성 후 LLM 호출
인터페이스 계층	FastAPI 기반 REST API 및 웹 UI, LMS·협업 도구 연동 지원



[그림 1] 시스템 아키텍처

[Fig 1] System Architecture

데이터 흐름은 ‘사용자 입력→쿼리 임베딩→벡터 검색→관련 맥락 수집→프롬프트 구성→LLM 생성→(선택) 후처리·검증→결과 반환’이며, [그림 2]에 정리하였다.



[그림 2] 번역 파이프라인 데이터 흐름

[Fig. 2] Translation Pipeline Data Flow

3.3 지식베이스 구축

지식베이스는 자료 수집, 텍스트 추출·정제, 의미 단위(또는 용어 항목 단위) 청킹, 임베딩, 메타 데이터 색인의 순서로 구축한다. 용어집은 ‘용어-정의-다국어 대응어-용례’ 구조로 관리한다. 이 구조는 검색 후 프롬프트 삽입 시 용어 일관성 제어에 유리하다.

3.4 번역 파이프라인

번역 파이프라인은 입력·전처리, 검색, 프롬프트 구성, 생성, 선택적 후처리·검증의 다섯 단계로 구성된다. 프롬프트에는 ‘제공된 용어 대응을 우선 사용할 것’이라는 명시적 제약을 포함한다. 본 시스템은 용어 일관성을 위해 생성 온도(temperature)를 낮게 설정한다.

3.5 용어 일관성 보장 메커니즘과 정적 삽입과의 관계

용어집 우선 검색, 프롬프트 제약, 동일 세션 내 이전 대응 재사용, 사후 검증 등을 통해 일관성을 유도한다. 중요한 점은 ‘전체 용어집을 프롬프트에 정적 삽입하는 방식’과의 차이이다. 용어집과 강의자료가 작을 때는 정적 삽입이 단순하고 효과적일 수 있다. 그러나 규모가 커지면 컨텍스트 길이 제한, 관련 없는 항목으로 인한 노이즈, 프롬프트 비용 증가가 발생한다. RAG는 필요한 맥락만 선별하여 이 문제를 완화하려는 접근이다. 따라서 평가에서는 정적 삽입 조건과의 비교가 필수적이다.

4. 프로토타입 구현

4.1 개발 환경

Python을 주 언어로 하며, 웹 서비스 계층은 FastAPI로 구현한다. 벡터 저장소는 오픈소스 벡터 데이터베이스를, 임베딩·생성은 다국어 지원 모델(상용 API 또는 로컬)을 사용하는 방식을 가정한다.

4.2 주요 모듈

프로토타입의 주요 모듈은 [표 3]과 같이 구성된다.

[표 3] 프로토타입 주요 모듈

[Table 3] Main Modules of the Prototype

모듈명	역할
ingestion	파일 업로드, 텍스트 추출, 청킹, 임베딩, 벡터DB 색인
retrieval	쿼리 임베딩, 유사도 검색, 하이브리드 검색, 결과 랭킹
prompt	검색 결과와 원문을 결합한 프롬프트 템플릿 관리
generation	LLM API 호출 및 응답 파싱
api	FastAPI 엔드포인트(문서 업로드, 번역 요청, 용어집 조회 등)

4.3 동작 예시

교수자가 용어집과 핵심 강의자료를 업로드하면 시스템이 색인한다. 학습자가 번역을 요청하면 관련 용어 정의·용례를 검색하여 LLM에 제공하고, 일관된 용어가 반영된 번역문과 참조 맥락을 반환한다. 구체적인 입출력 예시는 실험 단계에서 실제 데이터로 채운다.

5. 평가 설계

본 연구는 시스템 설계·프로토타입 논문으로, 대규모 실증 결과는 아직 포함하지 않는다. 본 절에서는 향후 수행할 평가의 설계와 방법론만을 제시한다.

5.1 비교 대상

[표 4]와 같이 네 가지 조건을 비교한다. 특히 ③과 ④의 대비가 본 연구의 핵심 질문(‘검색이 필요한가’)을 검증한다. 세 조건(②③④)에서 사용하는 LLM을 동일하게 통제하여, 검색 유무와 맥락 제공 방식의 효과를 분리 관찰한다. 용어집이 작을 때와 클 때를 나누어 실험하면 ‘규모에 따른 검색의 필요성’을 더 명확히 논증할 수 있다.

[표 4] 평가를 위한 비교 조건

[Table 4] Comparison Conditions for Evaluation

조건	설명	용어집·맥락 제공 방식
① 일반 MT	상용 신경망 기계번역 서비스	없음
② LLM 단독	동일 LLM에 원문·대상 언어만 제공	없음

③ 정적 삽입	LLM + 용어집 전체 프롬프트 삽입	용어집 전체 정적 포함(검색 없음)
④ RAG	제안 시스템 (검색 기반)	관련 맥락만 검색하여 동적 제공

5.2 평가 지표

5.2.1 번역 품질

자동 지표는 BLEU [11], chrF [12], COMET [13]를 사용하되, 소규모 인간 평가(유창성·의미 보존)를 병행한다. 비영어·전문용어·참조번역 확보가 어려운 경우 자동 지표의 한계가 크므로, 인간 평가의 비중을 높인다.

5.2.2 전공용어 일관성 비율

사전에 정의된 용어집 항목에 대해, 원문에 해당 용어가 등장한 경우 번역문에 약속된 대응어가 올바르게 사용되었는지를 비율로 측정한다.

조작적 정의 예: ‘용어 일관성 비율 = (올바르게 대응된 용어 출현 수) / (원문에 등장한 용어집 항목 총수)’.

형태소 변형·부분 일치 허용 여부, 대소문자·띄어쓰기 정규화 규칙을 사전에 확정해야 한다. 이 지표는 본 연구의 핵심 기여를 검증하는 데 직접 활용된다.

5.3 실험 절차

먼저 특정 COIL 교과의 강의자료·용어집으로 지식베이스를 구축한다. 다음으로 용어집 항목이 충분히 포함되도록 평가용 문장 집합을 구성한다. 이어 [표 4]의 네 조건으로 각각 번역을 생성하고, 자동 지표와 용어 일관성 비율을 계산한 뒤 소규모 인간 평가를 실시한다. 마지막으로 일관성이 깨진 사례와, RAG가 특히 도움이 된(또는 정적 삽입과 차이가 난) 사례를 정성 분석한다.

6. 결론 및 향후 연구

본 연구는 다국어 COIL 환경에서 전공용어 번역 불일치 문제에 대응하기 위해, 교과 맥락을 검색하여 LLM 번역에 반영하는 시스템을 설계하였다. 기술적으로는 기존 용어 제약 기계번역·용어집 주입 연구의 연장선에 있다. 본 연구의 기여는 다자·비영어권 COIL이라는 교육 맥락에 특화된 점과, 번역 품질과 함께 전공용어 일관성 비율을 평가 지표로 도입한 점이다.

평가 설계에서는 [표 4]와 같이 일반 MT, LLM 단독, 용어집 전체 정적 삽입, RAG의 네 조건을 비교하도록 구성하여, ‘검색이 단순 전체 삽입보다 나은가’라는 질문에 답할 수 있는 프레임에 제

시하였다. 현재 단계는 설계와 프로토타입에 초점을 맞춘 것으로, 실제 COIL 현장에서의 효과 검증은 향후 과제로 남는다. 본 연구는 시스템 설계와 평가 프레임워크 제시에 범위를 두며, 실증 검증은 후속 연구에서 다룬다.

향후 연구에서는 실제 다국어 COIL 수업에서 시스템을 적용하여 용어 일관성·학습자·교수자 경험을 측정하고, 용어집 규모에 따른 정적 삽입과 RAG의 성능·자연 트레이드오프를 분석하며, LMS·협업 도구와의 연동을 통해 현장 활용성을 강화할 계획이다.

References

- [1] J. Rubin, "Embedding collaborative online international learning (COIL) at higher education institutions," *Internationalisation of Higher Education*, vol. 2, pp. 27-44, 2017.
- [2] C. K. Y. Chan, "A scoping review of empirical research on Collaborative Online International Learning (COIL) implementation," *Educational Research Review*, vol. 49, Art. no. 100749, Nov. 2025, doi: 10.1016/j.edurev.2025.100749.
- [3] R. Fukkink, R. Helms, O. Spee, A. Mongelos, K. Bratland, and R. Pedersen, "Pedagogical dimensions and intercultural learning outcomes of COIL: A review of studies published between 2013-2022," *Journal of Studies in International Education*, vol. 28, no. 5, pp. 761-779, Nov. 2024, doi: 10.1177/10283153241262462.
- [4] P. Koehn and R. Knowles, "Six challenges for neural machine translation," in Proc. *First Workshop Neural Machine Translation*, Vancouver, Canada, Aug. 4, 2017, pp. 28-39, doi: 10.18653/v1/W17-3204.
- [5] M. Post and D. Vilar, "Fast lexically constrained decoding with dynamic beam allocation for neural machine translation," in Proc. *2018 Conf. North American Chapter Assoc. Comput. Linguistics: Human Language Technologies*, vol. 1, New Orleans, LA, USA, Jun. 1-6, 2018, pp. 1314-1324, doi: 10.18653/v1/N18-1119.
- [6] G. Dinu, P. Mathur, M. Federico, and Y. Al-Onaizan, "Training neural machine translation to apply terminology constraints," in Proc. *57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul. 28-Aug. 2, 2019, pp. 3063-3068, doi: 10.18653/v1/P19-1294.
- [7] K. Semenov, V. Zouhar, T. Koemi, D. Zhang, W. Zhou, and Y. E. Jiang, "Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies," in Proc. *8th Conf. Machine Translation (WMT)*, Singapore, Dec. 6-7, 2023, pp. 663-671, doi: 10.18653/v1/2023.wmt-1.54.
- [8] NLLB Team et al., "No Language Left Behind: Scaling human-centered machine translation," Jul. 2022, arXiv:2207.04672, doi: 10.48550/arXiv.2207.04672.
- [9] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, Online, Dec. 6-12, 2020, pp. 9459-9474.
- [10] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45-77, Winter 2007-8, doi: 10.2753/MIS0742-1222240302.
- [11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine

- translation," in Proc. *40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Philadelphia, PA, USA, Jul. 7-12, 2002, pp. 311-318, doi: 10.3115/1073083.1073135.
- [12] M. Popović, "chrF: Character n-gram F-score for automatic MT evaluation," in Proc. *10th Workshop Statistical Machine Translation (WMT)*, Lisbon, Portugal, Sep. 17-18, 2015, pp. 392-395, doi: 10.18653/v1/W15-3049.
- [13] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in Proc. *2020 Conf. Empirical Methods Natural Language Processing (EMNLP)*, Online, Nov. 16-20, 2020, pp. 2685-2702, doi: 10.18653/v1/2020.emnlp-main.213.