SEACrowd: 동남아시아 언어들을 위한 다국어 다중 모달 데이터 허브 및 벤치마크 제품군

홀리 로베니아[★,1,2] 라흐마드 마헨드라[★,3,2] 살사빌 마울라나 아크바르[★,2]
Lester James V. Miranda[★,4] Jennifer Santoso[★,5] Elyanah Aco[★,(6)] 아크단 파딜라(★) (,) (7) 조니벡 만수로프(★) (,) (8) 조셉 마빈 임페리얼(★) (,) (9) (,) (10) 온노 P. 캄프만(★) (,) (11)
Joel Ruben Antony Moniz[★,6] Muhammad Ravi Shulthan Habibi[★,3,2] Frederikus Hudi[★,12,(13)] 레이리 몬탈란(★) (,) (1) 라이언 이그나티우스(6) 조아니토 아길리 로포(14) 윌리엄 닉슨(15)
보예 F. 칼슨[16] 제임스 자야[6] 랴안디토 디안다루[6] 가오 유제[6] 패트릭 아마데우스[15] 왕 빈[6] 얀 크리스티안 블레이즈 크루즈[8,17] 천시 화 이트하우스[18] 이반 할림 파르모낭간[19]
마리아 켈리[15] 웬위 장[6] 럭키 수산토[20] 레이너드 아다 룽다[21] 소니 라주아르디 헤르마완[22] 댄 존 벨라스코[17] 무함마드 데한 알 카우츠아르[15]
윌리 피트라 헨드리아[6] 야스민 모슬렘[23] 노아 플린[24] 무함마드 파리드 아딜라주아르다[8] 리 하오천[6] 요하네스 리[15] R. 다만후리[25] 쑨 슈오[6] 무함마드 레자 코리브[26]아미르베크 자니베코프[8] 웨이 치 룽[1] 콴 V. 도[27] 니클라스 뮌니호프[28]
탄라다 판수완[18] 일함 피르다우시 푸트라[6] 얀 쉬[29,27] 응이 치아 타이[1]
아유 푸르와리안티[6,30] 세바스찬 루더[31] 윌리엄 치[1] 피랏 림콘초티왓[★,32] 알함 피크리 아지[★,(8)] 세드릭 케(★) (,) (33) 겐타 인드라 위나타(★) (,) (35) (,) (2) 루오천 장(★) (,)(34) 파즈리 코토(★) (,) (8) (,) (2) 정신 용(★) (,) (34) 사무엘 카야위자야(★) (,) (31) (,) (27) (,) (2)

[1] AI Singapore [2] IndoNLP [3] 인도네시아 대학 [4] 앨런 인공 지능 연구소 [5] RevComm, Inc.
[6] 독립 연구원 [7] 도호쿠 대학 [8] MBZUAI [9] 바스 대학교 [10] 필리핀 국립대학교
[11] 보건의료혁신처(MOHT) [12] 나이스츠(NAIST)(13) 워크스 애플리케이션스 랩(Works Applications Lab)(14) 가자마다 대학교(Universitas Gadjah Mada)(15) 반둥 공과대학교(Institut Teknologi Bandung)[16] 베이징 인공지능 연구소(BAAI) (17) 삼성리서치 필리핀 (18) 케임브리지 대학교 (19) 퀸즐랜드 공과대학교 (20) 모나쉬 대학교 인도네시아 (21) 임페리얼 칼리지 런던 (22) 독립 디자인 엔지니어 (23) 베링 랩 (24) 아마존 (25) 디포네고로 대학교 (26) NUS (27) HKUST (28) Contextual AI (29) Huawei Noah's Ark Lab (30) Prosa.ai (31) Cohere (32) VISTEC (33) Toyota Research Institute (34) Brown University (35) Capital One
· 주요 기여자

초록

동남아시아(SEA)는 1,300개 이상의 토착 언어와 6억 7,100만 명의 인구를 보유한 풍부한 언어적 다양성과 문화적 다채로움을 특징으로 하는 지역이다. 그러나 동남아시아 언어에 대한 현대 AI 모델의 성능은 동남아시아 출처의 텍스트, 이미지, 청각 데이터 세트의 심각한 부족으로 인해 저하된다. SEA 언어에 대한 모델 평가가 어려운 이유는 고품질 데이터 세트의 부족과 더불어 영어 훈련 데이터가 주를 이루기 때문이며, 이는 잠재적인 문화적 오표현에 대한 우려를 제기합니다. 이러한 과제를 해결하기 위해, 우리는 SEACrowd라는 공동 이니셔티브를 소개합니다. 이 이니셔티브는 세 가지 양식(문자, 음성·영상)에 걸쳐 거의 1,000개의 동남아시아 언어로 표준화된 코퍼스와 벤치마크를 제공함으로써 자원 격차를 해소하기 위한 포괄적인 자원 허브(1)를 통합합니다. 우리는 SEACrowd에 포함된 13가지 작업에 걸쳐 36개 토착어로 AI 모델의 성능을 평가하여 동남아시아의 현재 AI 환경에 대한 귀중한 통찰력을 제공합니다. 또한, 우리는 촉진하기 위한 전략을 제안합니다. 동남아시아 AI의 미래를 위한 잠재적 유용성과 자원 형평성을 극대화하는 더 큰 AI 발전을 추구합니다.

1 서론

동남아시아(SEA)에는 1,300개의 토착 언어(전 세계 언어의 18%)와 6억 7,100만 명(전 세계 인구의 8.75%)이 거주함에도 불구하고, 이 지역의 텍스트, 이미지, 오디오 데이터셋은 머신러닝 모델에서 현저히 부족하게 반영되고 있다. 이러한 결핍은 SEA 언어에 대한 모델 품질에 부정적인 영향을 미친다. 두 가지 대표적인 사전 훈련 자원인 Common Crawl(3)과 C4(Xue et al., 2021)에서 동남아시아 언어의 커버리지는 극히 제한적입니다. 각각 2.36%(11개 언어)와 10.62%(11개 언어)에 불과합니다. 텍스트를 넘어선 모달리티에서는 표현이 더욱 제한적입니다. 예를 들어, 가장 큰 다국어 음성 코퍼스 중 하나인 Common Voice에는 6개의 동남아시아 원주민 언어가 포함되어 있습니다(Conneau et al., 2021; Ardila et al., 2020). 또한 가장 큰 다국어 시각-언어 코퍼스 중 하나인 LAION-5B에는

[1] https://seacrowd.github.io/seacrowd-catalogue/
[2] https://github.com/SEACrowd/seacrowd-datahub/
[3] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

2024년 자연어 처리 경험적 방법론 학회 논문집, 5155–5203쪽 2024년 11월 12-16일 ©2024 계산언어학회

(VL) 코퍼스는 12개의 동남아시아 원주민 언어를 포함한다 (Schuhmann et al., 2022). 다른 동남아시아 원주민 언어에 대한 데이터 세트도 존재하지만, 종종 흩어져 있거나 문서화가 불충분하거나 품질과 형식이 다양하여 접근과 사용이 어렵습니다 (Cahyawijaya 외, 2023a; Joshi 외, 2020; Aji 외, 2023).

평가 측면에서, 이러한 언어들에 대한 고품질 테스트 세트의 부족은 동남아시아 언어 모델 평가를 더욱 복잡하게 만든다. 동남아시아 지역에 1,300개 이상의 언어가 존재함에도 불구하고, 기존 연구들(Winata et al., 2023; Cahyaw-ijaya et al., 2021; Koto and Koto, 2020; Zhang et al., 2024; Wang et al., 2024; Nguyen et al., 2023; Leong et al., 2023)은 총 10개 미만의 동남아시아 언어만을 평가했습니다. 현재 모델들의 대부분의 동남아시아 언어에 대한 실제 성능은 여전히 대부분 알려지지 않았다. 더욱이, 영어 중심의 훈련 데이터의 지배력은 대표성이 부족한 동남아시아 언어들로 텍스트, 이미지 또는 오디오를 생성할 때 문화적 편향을 초래할 수 있다(Søgaard, 2022; Talat et al., 2022). 또한 Durmus et al. (2023); AlKhamissi et al. (2024); Cahyawijaya et al. (2024a)은 대규모 언어 모델(LLMs)에서 학습된 표현이 동남아시아의 지역적 문화적 가치를 반영하지 못하는 경우가 많다는 점을 보여주었다(Koto et al., 2024; Liu et al., 2024; Adilazuarda et al., 2024). 이는 현재의 대규모 언어 모델이 이 지역을 위한 자연스럽고 고품질의 텍스트를 생성할 수 있는 능력에 대한 우려를 제기합니다. 또한 언어 지원의 불균형은 기술 접근에 언어 장벽을 만들고 해당 언어를 사용하지 않는 소수 집단을 소외시킬 위험이 있습니다. 주류 언어.

본 연구에서는 자원, 평가, 생성 품질이라는 과제를 해결함으로써 동남아시아 언어에 대한 현재 AI 진전을 조사합니다. 우리의 기여는 세 가지 측면으로 이루어집니다:

- 우리는 SEACrowd라는 포괄적이고 표준화된 자원 센터를 통해 약 500개의 코퍼스를 중앙화하고 표준화함으로써 자원 격차를 해소합니다. 이 센터는 텍스트, 이미지, 오디오 세 가지 모달리티를 아우릅니다.

  텍스트, 이미지, 오디오 세 가지 모달리티를 아우르는 포괄적이고 표준화된 자원 센터입니다.
- 또한 SEACrowd 벤치마크를 통해 동남아시아 언어 평가 격차를 해소하였습니다. 이 벤치마크는 38개 동남아시아 원주민 언어를 대상으로 3가지 모달리티(텍스트, 이미지, 오디오)에 걸쳐 13가지 과제를 다루며, 다양한 스펙트럼의 AI 모델 성능에 대한 통찰력을 제공합니다. 더 나아가 본 연구는 9개 동남아시아 언어에서

  기존 대규모 언어 모델의 생성 출력은
  "번역체"에 더 가깝게
  9개 동남아시아 언어에서 자연어 데이터보다 더 많은 데이터를 확보했습니다.
- 동남아시아 AI의 향후 발전을 위한 통찰력과 전략을 제시합니다.

2 SEACrowd

SEACrowd는 동남아시아 지역을 중심으로 활동하는 연구자 및 엔지니어들의 협력으로 개발된 동남아시아 최초의 포괄적인 AI 데이터셋 수집 프로젝트입니다. 1절에서 언급한 바와 같이, 동남아시아에서는 자원 부족과 데이터의 분산성이 중요한 과제입니다. SEACrowd는 두 가지 주요 기여를 통해 이러한 문제를 해결합니다: 1) **데이터 시트를 통합하여** 데이터 검색성을 향상시키고,
2) **데이터 로더를 표준화하여** 사용 편의성을 높이고, 특히 다중 데이터셋 로딩 시 효율성을 개선합니다. 또한 데이터 출처 추적 관행(Longpre et al., 2023)을 준수하여 데이터셋 소유자의 독점적 권리를 보호합니다.

데이터시트 통합 SEA 언어 및/또는 문화권에서 텍스트, 오디오, 이미지를 포함한 모든 양식의 공개 데이터 세트에 대해 기여자들에게 데이터시트 양식(Gebru 외, 2021)을 제출하도록 요청했습니다. 이 데이터 시트에는 데이터 하위 집합, 설명, 작업, 언어, 라이선스, URL 액세스, 주석 방법, 주석 유효성 검사, 관련 출판물, 출판 장소 및 데이터 분할과 같은 각 데이터 세트에 대한 자세한 정보가 포함됩니다. 제출된 각 데이터 시트는 정확성을 보장하기 위해 수동으로 확인하고 필요한 경우 수정합니다.

데이터 로더 표준화 승인된 각 데이터시트에 대해, 통합 데이터 소스의 38.4%만이 원래 Hugging Face[4]에 호스팅되었기 때문에즉시 사용 가능한 데이터 접근을 용이하게 하기 위해 표준화된 데이터 로더 래퍼를 생성했습니다. 다양한 작업 유형을 지원하기 위해, 서로 다른 데이터 구조와 모달리티를 지원하도록 표준화된 seacrowd 스키마를 신중하게 설계했습니다(부록 F 참조). 또한 데이터 출처 관행(Long-pre 외, 2023)을 준수하고 데이터 로더에 관련 메타데이터(예: 라이선스)를 문서화했습니다. 더불어 데이터 소유자와 협력하여 세 개의 비공개 데이터셋을 공개 데이터셋으로 전환하는 데 성공했습니다.

이러한 노력의 결과로 SEACrowd 카탈로그에는 498개의 데이터 시트가, SEACrowd 데이터 허브(§2.1)에는 399개의 데이터 로더가 구축되었습니다. 특히, 중앙 집중식 데이터 저장소는 약 1,000개의

SEACrowd Data Hub (§2.1)에 399개의 데이터 로더를 구축하는 성과를 거두었습니다. 특히, 우리의 중앙 집중식 데이터 저장소는 약 1,000개의 SEA 언어를 포괄하여 SEACrowd가 포착한 광범위한 언어적 다양성을 강조합니다.
이 점은 SEACrowd가 포착한 방대한 언어적 다양성을 보여줍니다. 이에 대해 자세히 설명합니다.

[4] https://huggingface.co/

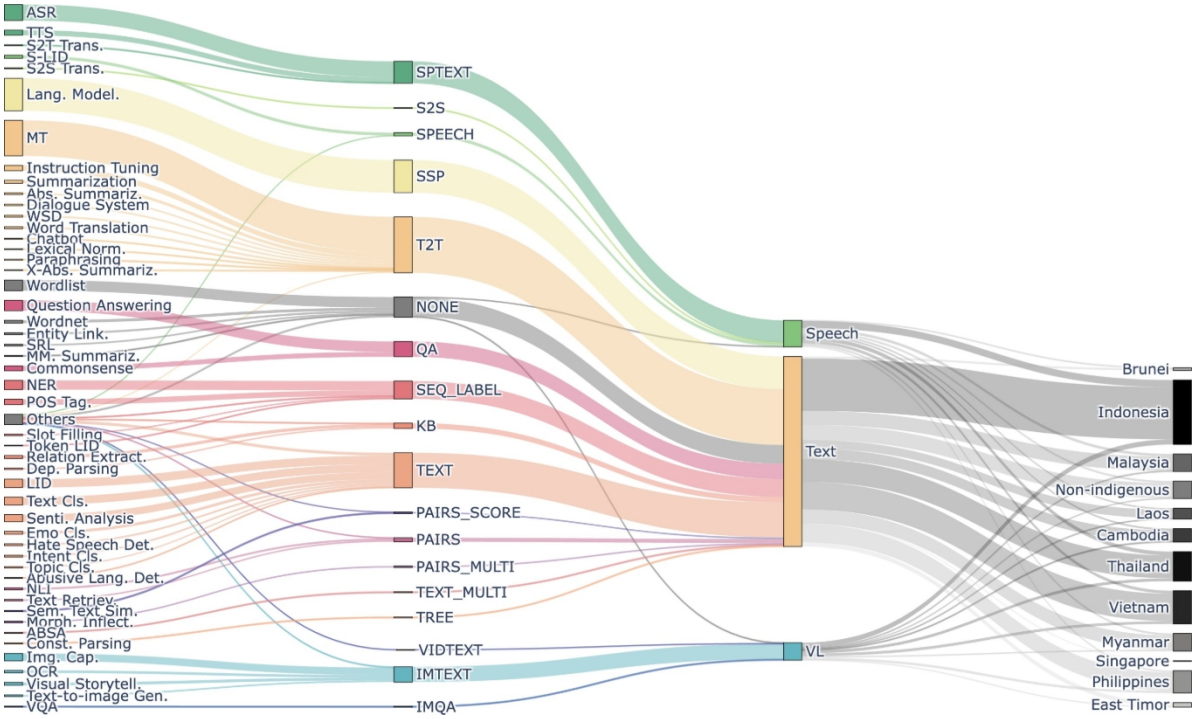


그림 1: SEACrowd의 498개 데이터시트에 걸쳐 작업, 스키마, 모달리티 및 언어 영역 간의 매핑.

§2.2의 SEACrowd 데이터셋 통계. SEACrowd의 기여 가이드라인, 진행 세부사항 및 검토 절차는 부록 C, D, E에 있습니다.

2.1 SEACrowd 카탈로그 및 데이터 허브

SEACrowd는 상호 연결된 두 플랫폼으로 구성됩니다: SEACrowd 카탈로그[5]와 SEACrowd 데이터 허브입니다. 이 플랫폼들은 데이터시트 제출을 통합하고 SEACrowd를 위한 표준화된 파이프라인을 제공하기 위해 함께 작동합니다. 구체적으로, 카탈로그는 데이터시트(메타데이터)를 보관하는 반면, 데이터 허브는 스키마 및 구성(부록 F)을 위한 표준화된 데이터 로더와 seacrowd 라이브러리([6])를 저장합니다. 이 시스템들은 데이터시트와 데이터 로더에 대한 정보를 공유하여 사용자가 이를 원활하게 탐색하고 활용할 수 있도록 합니다.

2.2 SEACrowd의 데이터셋

SEACrowd는 SEA 언어의 다양한 작업을 포함한 498개의 데이터 시트를 통합하고, 데이터 로더를 통해 그 중 399개에 대한 표준화된 접근을 제공합니다. 그림 1에서 볼 수 있듯이, 약 81%의

SEACrowd의 데이터셋은 텍스트 데이터이며, 나머지 약 8%와 약 11%는 각각 시각적 언어(VL) 및 음성 데이터입니다. SEACrowd가 다루는 SEA 원주민 언어의 전체 목록과 해당 SEA 지역과의 대응 관계는 아래와 같습니다.

언어 목록과 해당 동남아시아 지역 매핑은 부록 K. 데이터 세트의 약 53%는 상업적으로 허용적인 라이선스를 가지고 있습니다.

SEACrowd에는 총 83개의 작업이 제공되며, 그 내역은 NLP 66개(예: 욕설 감지, 의도 분류, 지시문 조정, 명명된 엔티티 인식 등), VL 10개(이미지-텍스트 생성, 수화 인식, 비디오 캡션, 등), 음성(예: 자동 음성 인식, 텍스트 음성 변환, 음성 감정 인식 등) 7개로 구성됩니다. 이 작업들은 부록 F에 설명된 20개의 데이터 로더 스키마로 표준화됩니다. SEACrowd의 리소스에 대한 추가 논의는 §5.1에 있습니다.

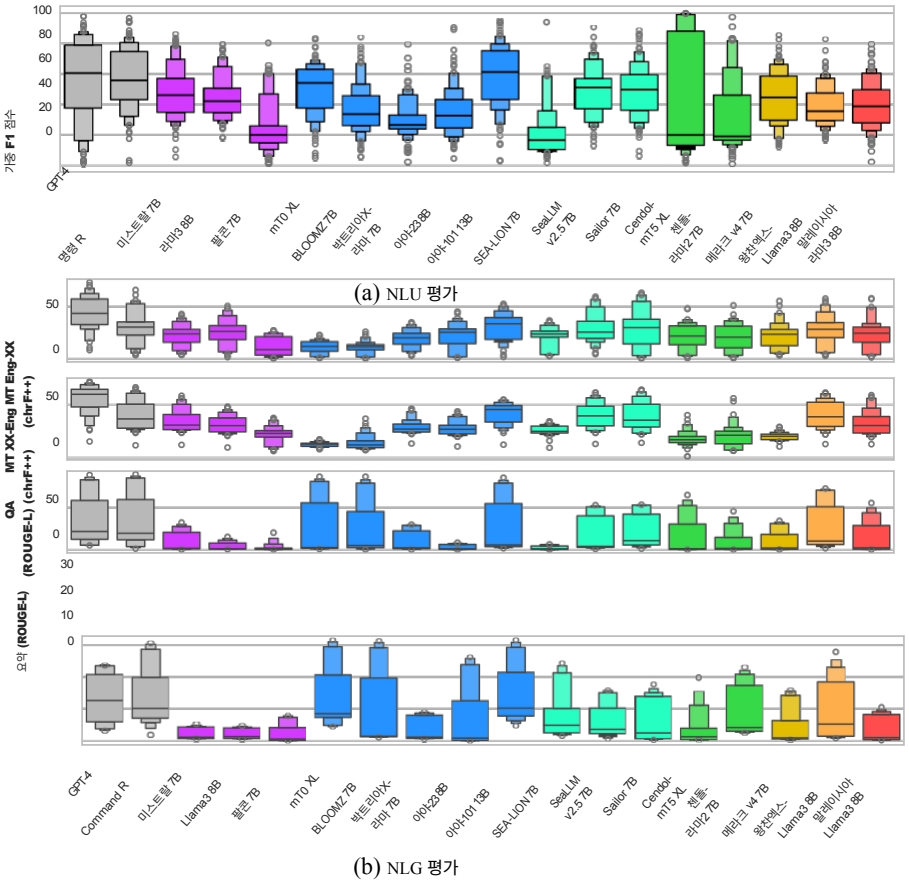
3 SEACrowd 벤치마크

최첨단 모델의 성능을 이해하기 위해, 다양한 아키텍처와 훈련 접근법을 활용한 기존 대규모 언어 모델(LLM), 대규모 영상 모델(VLM), 음성 모델에 대한 포괄적인 평가를 수행합니다. 벤치마크 세트[7]를 구축하기 위해, §2.2에 제시된 데이터 중 수동으로 주석 처리 및/또는 검증된 데이터의 하위 집합을 선택합니다. 평가에 사용된 데이터 하위 집합, 기준 모델 및 프롬프트에 대한 자세한 내용은 부록 G.1, G.2 및 G.3에 제시되어 있습니다.

[5] SEACrowd 카탈로그는 csv 형식으로도 제공됩니다.

[6] 모든 코드는 Apache License 2.0에 따라 이용 가능합니다.

[7] https://github.com/SEACrowd/seacrowd-experiments



(a) NLU 평가

(b) NLG 평가

그림 2: 동남아시아 언어에서 NLU 및 NLG 작업에 걸친 제로샷 모델 성능.

모델	지니 ↓
상업용	
GPT-4	0.155
Command-R	0.184
영어	
미스트랄	0.159
Llama3	0.131
팔콘	0.238
다국어	
mT0	0.131
BLOOMZ	0.228
BactrianX-Llama	0.163
AYA-23	0.183
AYA-101	**0.095**
SEA 지역	
SEA-LION	0.204
SeaLLM v2.5	0.116
Sailor	0.145
SEA 국가	
Cendol-mT5	0.378
Cendol-Llama2	0.267
메라크 v4	0.199
왕찬X-라마3	0.153
말레이시아 Llama3	0.179

표 1: 지니 계수 가중치 기반 기준선 간 언어 평등도 인구별 ($\tau = 0.5$).

3.1 데이터 세트

NLP 당사의 자연어 이해(NLU) 벤치마크는 131개의 데이터 하위 집합과 7가지 작업으로 구성됩니다: 감성 분석, 주제 분류, 자연어 추론(NLI), 상식 추론, 시험형 객관식 질문 답안(QA), 문화 이해, 독해력. 영어(ENG)와 33개의 동남아시아 토착 언어를 다룹니다.

우리는 자연어 생성(NLG) 벤치마크에 100개의 데이터 하위 집합을 활용하며, 이는 영어와 동남아시아 언어 간의 양방향 기계 번역(MT), 요약, 추출형 또는 요약형 질문 응답을 포함하며, 27개의 동남아시아 토착 언어를 다룹니다.

음성 15개 동남아시아 토착 언어에서 음성 모델의 능력을 평가하기 위해 19개의 자동 음성 인식(ASR) 데이터 하위 집합을 사용합니다.

VL 우리는 필리핀어(FIL), 인도네시아어(IND), 태국어(THA), 베트남어(VIE) 등 4개 동남아시아 원주민 언어로 된 네 가지 데이터 하위 집합을 사용하여 이미지 캡셔닝 모델을 평가합니다. 이러한 평가 규모 차이는 SEACrowd 내 데이터셋 중 극소수만이 VL 데이터셋이며, 그중에서도 인간에 의해 주석이 달린 데이터셋은 더욱 적기 때문입니다. 데이터셋만이 VL 데이터셋이며, 그중에서도 인간이 주석을 단 데이터셋은 더 적기 때문입니다.

3.2 기준선

모델 아키텍처, 모델 규모, 사용된 언어, 관련 논문 및 기타 측면에 대한 자세한 내용은 부록 G.2를 참조하십시오.

NLP 동남아시아 언어에 대한 지시어 튜닝된 대규모 언어 모델의 제로샷 성능을 평가하기 위해, 우리는 두 개의 상용 모델(즉, GPT-4(OpenAI et al., 2024) 및 Command-R[8])과 17개의 오픈소스
기준 모델을 벤치마킹했습니다. 대부분은 약 70억~130억 매개변수를 가집니다. 오픈소스 모델은 다음과 같이 분류합니다.
언어별 사전 훈련 및/또는 지시 튜닝 적용 범위에 따른 기준 모델: 1) **영어** : Llama3 (Touvron et al., 2023), Mistral (Jiang et al., 2023), Falcon (Almazrouei et al., 2023); 2) **다국어:** AYA-101, AYA-23 (Üstün et al., 2024), mT0, BLOOMZ (Muennighoff et al., 2022), BactrianX-Llama (Li et al., 2023a);
3) **동남아시아 지역** 기반: SEA-LION (Singapore, 2023), Sailor (Dou et al., 2024), SeaLLM (Nguyen et al., 2023); 및 4) **동남아시아 국가별**:

[8] https://docs.cohere.com/docs/command-r

인도네시아의 Cendol-mT5, Cendol-Llama2 (Cahyawijaya 외, 2024b), Merak (Ichsan, 2023), 태국의 WangchanX-Llama3 (Phatthiyaphaibun 외, 2024)·말레이시아의 Malaysian-Llama3[9].

음성 우리는 동남아시아 언어의 음성 전사를 위한 최첨단 **다국어 사전 훈련된** 음성 모델의 제로샷 성능을 평가합니다. 구체적으로, 정확한 전사를 보여준 Whisper v3 (Radford et al., 2023), MMS 1B (Pratap et al., 2024), Seamless M4T v2 (Communication et al., 2023)를 고려합니다.

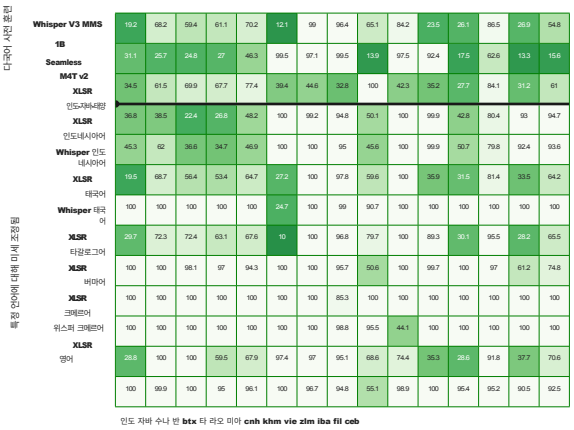


그림 3: 기존
동남아시아 언어의 기존 ASR 작업에 대한 음성 모델 오류율(%↓).

다중 언어에 대한 미세 조정 없이 설명합니다. 또한 **특정 언어(**SEA 또는 영어)**에 미세 조정된** 모델을 포함하며, 이는 1) Wav2Vec2 XLSR(Conneau et al., 2021) 및 2) XLS-R(Babu et al., 2021)을 기반으로 특정 언어(SEA 또는 영어)에 대해 미세 조정된 모델을 포함합니다. 이 모델들은 다양한 언어의 원시 음성 파형을 사전 훈련하여 다국어 음성 표현 학습으로 잘 알려져 있으며, XLS-R은 더 광범위한 언어 커버리지를 제공합니다. 또한 3) 다양한 언어의 음성 스펙트로그램에 대한 약한 지도 사전 훈련을 활용하는 Whisper도 포함됩니다. 구체적인 미세 조정 모델은 다음과 같이 평가되었습니다: XLSR은 IND, JAV, SUN에서; XLSR과 Whisper는 인도네시아어(IND)에서; XLSR과 Whisper는 태국어(THA)에서; XLS-R은 타갈로그어(TGL)에서; XLS-R은 버마어(MYA)에서; 크메르어(KHM)에 대해 XLS-R과 Whisper; 영어(ENG)에 대해 XLSR. 자세한 내용은 부록 G.2를 참조하십시오.

VL 우리는 주로 **영어** 사전 훈련 및 지시 따르기 데이터로 훈련된 최첨단 VLM을 고려한다: LLaVA (Liu et al., 2023b,a), In-structBLIP (Dai et al., 2024), Idefics2 (Lau-rençon et al., 2024), 그리고 **다국어** 방식으로 훈련된 VLM들: mBLIP (Geigle et al., 2023) 및 PaliGemma (Gemma Team et al., 2024)를 평가하여 동남아시아 언어에서의 이미지 캡션 생성 능력을 평가한다.

3.3 실험 설정

모든 평가는 제로샷 방식으로 수행합니다. 각 NLU 작업에 대해 영어로 된 3개의프롬프트 템플릿을, 각 NLG 작업에 대해 1개의 프롬프트 템플릿을 사용합니다. NLU 작업에서는 가중 F1 점수를, NLG 작업에서는 n-그램 참조 기반 지표인 chrF++(Popovic´, 2015, 2017) 및 ROUGE-L(Lin, 2004)을 모델 성능 측정 기준으로 활용합니다. VL의 경우, 영어 프롬프트 템플릿 외에도 데이터 하위 집합별로 해당 동남아시아 원주민 언어로 된 프롬프트 템플릿을 사용합니다. 우리는 보고합니다.

[9] https://huggingface.co/mesolitica/malaysian-llama-3-8b-instruct-16k

이미지 캡션 작업에 대한 CIDEr (Vedantam et al., 2015). ASR의 경우, 라틴 문자를 사용하는 언어에는 단어 오류율(WER)을, 비라틴 문자를 사용하는 언어에는 문자 오류율(CER)을 사용합니다.

4 결과 및 분석

4.1 SEA 언어에 대한 최첨단 모델

LLM 그림 2a와 2b는 NLU 작업과 NLG 작업 모두에서 동남아시아 언어에 대한 LLM 기준 모델의 전반적인 성능을 보여줍니다. 우리의 NLU 평가에서, 101개 언어를 포괄하는 대규모 다국어 명령어 튜닝 언어 모델인 AYA-101이 최고의 제로샷 성능을 보였습니다. 그 다음으로는 상용 기반 모델이 뒤를 이었습니다.
라인들은 가중 F1 점수 중앙값이 약 0.6을 달성합니다. 특히 SEA 언어 모델로 훈련된 Sailor와 SeaLLM 모델들도 경쟁력 있는 성능을 보입니다. 마찬가지로 mT0는 강력한 성능을 보입니다.
SEA 언어로 훈련된 모델들도 경쟁력 있는 성능을 보여줍니다. 마찬가지로 mT0는
일반화 능력을 보입니다. 이는 사전 훈련 과정에서 약 100개 언어(동남아시아 언어 포함)에 노출되었기 때문입니다.
동남아시아 지역(Muennighoff et al., 2022). 반면 대부분의 영어 및 동남아시아 국가별 기준 모델은 상대적으로 낮은 성능을 보였는데, 이는 영어에 대한 좁은 초점이나 Cendol의 인도네시아어, WangchanX-Llama3의 태국어와 같이 제한된 동남아시아 언어 집합에 기인한 것으로 추정된다. 기계번역(MT) 작업에서도 유사하고 일관된 경향이 관찰되지만, 추상적/추출적 QA 및 요약 작업에서 기준 모델들의 낮은 점수는 이러한 작업에서 동남아시아 언어로 수용 가능한 출력을 생성하는 데 효과적이지 않음을 나타내며, 특히 오픈소스 기준 모델에서 두드러집니다. 부록 G.4에는 언어별 대규모 언어 모델(LLM)의 성능이 설명되어 있습니다.

동남아시아 언어 간 모델 성능의 평등성을 분석하기 위해 Khanuja et al. (2023)을 따라 지니 계수를 활용합니다. 이 계수는 원래

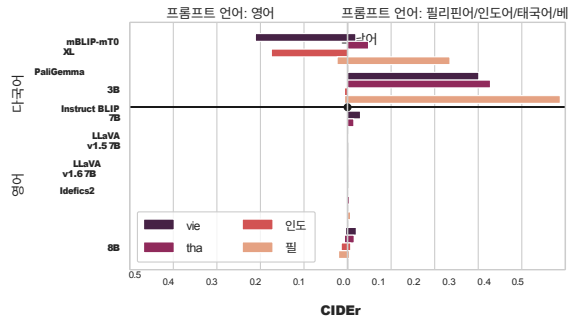


그림 4: 기존 VLM은 동남아시아 언어에서 수준 이하의 이미지 캡션을 생성합니다. 우리는 소득 평등을 관찰하기 위해 사용된 CIDEr(Vedantam et al., 2015)를 보고합니다.

수요에 가중치를 부여하고 τ로 매개변수화한 소득 평등 관측에 사용됨 (Dorfman, 1979). 여기서 $\tau = 1$은 인구통계학적 개념의

언어 인구 규모를 고려한 수요를 나타내며, $\tau = 0$은 인구 규모를 고려하지 않습니다(Blasi et al., 2022). 표 1은 다국어 및 동남아시아 지역 기반 모델과 같이 더 많은 동남아시아 언어로 훈련된 모델이 일반적으로 더 높은 언어 평등성을 보임을 보여줍니다. 예를 들어, Command-R과 GPT-4는 성능 면에서 AYA-101 및 mT0과 경쟁력이 있지만, AYA-101과 mT0은 연구 대상인 모든 동남아시아 언어에서 더 높은 평등성을 보여줍니다. 이 경향은 다양한 τ 값에서도 일관되게 나타납니다(부록 G.5 참조).

음성 모델 그림 3은 동남아시아(SEA) 지역 언어별 자동 음성 인식(ASR)에서 상용 음성 모델의 성능을 오류율(%)로 측정한 결과를 제시합니다. 본 음성 평가에 포함된 15개 SEA 언어 중 9개는 오스트로네시아어족에 속합니다. 나머지 6개 언어는 오스트로아시아어족에 속하는 KHM(크메르어)과 VIE(베트남어), 중국티베트어족에 속하는 CNH(중국어)와 MYA(몽골어), 크라다이어족에 속하는 THA(태국어)와 VIE(베트남어)입니다. 다국어 사전 훈련된 기준 모델들은 언어별로 차이는 있으나 전반적으로 경쟁력 있는 일반화 능력을 보입니다. 예를 들어 Whisper v3는 인도네시아어(IND), 말레이어(ZLM), 필리핀어(FIL), 태국어(THA), 베트남어(VIE)와 같은 국가 언어에서 현저히 높은 효과를 보인 반면, 다른 토착 언어에서는 최적의 성능을 발휘하지 못했습니다. 반면 Seamless M4T v2는 언어 전반에 걸쳐 보다 균형 잡힌 성능을 보였습니다. 파인 튜닝된 베이스라인의 경우, 학습된 언어에서 오류율이 감소합니다. 그러나 파인 튜닝된 Whisper 모델은 Wav2Vec2 XLSR 및 XLS-R 대비 대상 언어에 대한 최적화를 더 잘 수행하면서도 다른 동남아시아 언어에서의 원래 성능을 유지합니다. 두 모델 모두 다국어 환경에서 사전 훈련되었음에도 불구하고 말입니다.

모델	자연어 출력
SEA-LION	**58.57%**
아야-23	43.57%
Sailor	37.86%
첸돌-라마2	37.37%
말레이시아 라마3 왕찬X-라마3 30.24	36.90%
팔콘	29.52%
BactrianX-Llama	28.10%
SeaLLM	27.38%
Merak	26.19%
BLOOMZ	25.00%
Cendol-MT5	24.05%
커맨드-R	20.95%
mT0-XL	19.76%
미스트랄	19.52%
GPT-4	16.67%
Llama3	14.05%
AYA-101	8.33%

(a) 모델별 평균

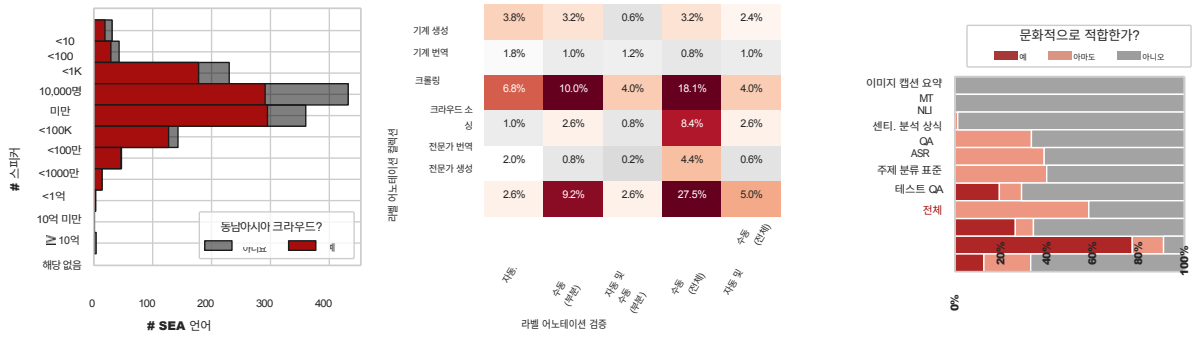
언어	자연어 출력
인도네시아어 (IND)	**41.58%**
베트남어 (VIE)	37.31%
태국 (THA)	34.21%
크메르어 (KHM)	29.21%
라오어 (LAO)	28.42%
말레이 (ZLM)	22.24%
버마어 (MYA)	19.47%
필리핀 (FIL)	12.22%
영어 (ENG)†	8.95%

(b) 언어별 평균

표 2: 현재의 대규모 언어 모델(LLM)은 동남아시아 언어(SEA)로 자연스러운 텍스트를 생성하는 데 여전히 한계가 있다. †전 세계적으로 통용되는 것이 아닌 동남아시아 지역에서 사용되는 방언을 기준으로 함.

이 관찰 결과는 Rouditchenko 등(2023)의 연구 결과와 일치한다. 해당 연구에서는 사전 훈련 과정에서 언어 및 언어군별로 노출된 시간 수가 모델 간 성능 비교를 예측하는 지표가 된다는 점을 발견했으며, Whisper의 사전 훈련 데이터 노출 시간이 XLSR보다 이 네 언어군에서 더 길었다.

VLMs 그림 4는 동남아시아 원주민 언어의 이미지 캡션 작성에 대한 상용 VLMs의 제로샷 성능을 보여줍니다. LLM의 제로샷 교차언어 일반화 능력(Huang et al., 2021; Täckström et al., 2012; Neubig and Hu, 2018; Artetxe et al., 2020), 영어로만 훈련된 VLM(즉, InstructBLIP, LLaVA, Idefics2)은 이러한 능력을 보여주지 못하며 동남아시아 언어로 적절한 이미지 캡션을 생성하는 데 어려움을 겪습니다. 다국어 VL 사전 훈련은 정렬된 다국어 표현을 달성하는 데 핵심적입니다(Burns et al., 2020; Li et al., 2023b; Huang et al., 2021). 예를 들어, PaliGemma와 mBLIP은 관련 동남아시아 언어로 프롬프트를 입력할 때 태국어(THA)와 필리핀어(FIL)에서 더 나은 이미지 캡션을 생성합니다. 그러나 영어로 프롬프트를 입력할 경우, 이러한 다국어 기반 모델들의 성능은 현저히 달라집니다. PaliGemma의 성능은 완전히 붕괴되는 반면, mBLIP의 성능은 다양한 동남아시아 언어에 따라 증가하거나 감소하는 양상을 보입니다. 이는 다국어 VLM이 지시문과 작업에 사용된 서로 다른 언어 전반에 걸쳐 일관된 성능을 유지할 수 있는지 의문을 제기합니다. 이는 다국어 VL 모델의 성능이 언어 간에 어떻게 달라지는지, 그리고 이러한 성능 차이가 모델의 다국어 처리 능력과 관련이 있는지 여부를 명확히 할 필요성을 강조합니다.



(a) 언어 지원 범위 (b) 주석 품질 (c) 문화적 적합성

그림 5: 언어 지원 범위, 주석 품질 및 문화적 적합성 측면에서 SEA의 자원 격차.

이러한 변이를 주도하는 메커니즘과 다양한 언어적 맥락에서 VLM의 견고한 다국어 성능을 달성하는 방법에 대한 추가 연구가 필요하다. 이러한 역학 관계를 이해하는 것은 VLM의 일반화 능력을 향상시키고 모든 언어에서 공평한 성능을 보장하는 데 매우 중요합니다. 대부분의 관련 연구는 단일 언어 시각적 지시 조정(Liu et al., 2023b; Gong et al., 2023; Zhu et al., 2024)에 초점을 맞추고 있음에도 불구하고 말입니다.

4.2 동남아시아 언어에서의 생성 품질: 번역체 대 자연어

동남아시아 언어에서의 번역어 분류 동남아시아 언어에서 대규모 언어 모델의 생성 품질을 분석하기 위해, 우리는 번역어와 자연어 텍스트를 구분하는 텍스트 분류기를 구축하였다(Riley et al., 2020). 영어(ENG)와 8개 동남아시아 언어(인도네시아어(IND), 크메르어(KHM), 라오어(LAO), 버마어(MYA), 필리핀어(FIL), 태국어(THA), 베트남어(VIE), 말레이어(ZLM))로 구성된 약 39.9k 및 51.5k 문장을 포함하는 49개 및 62개 데이터 하위 집합을 사용하여 번역어 분류 훈련 및 테스트 데이터셋을 구축했습니다. 인도네시아어(IND), 크메르어(KHM), 라오스어(LAO), 버마어(MYA), 필리핀어(FIL), 태국어(THA), 베트남어(VIE), 말레이어(ZLM). 훈련 및 테스트 데이터는 부록 H.1에 상세히 기술되어 있습니다.

우리는 이러한 데이터를 사용하여 mDeBER-TaV3(He et al., 2020, 2022)[10]의 분류기를 미세 조정하고, 9개 언어에 걸쳐 번역체(translationese)를 예측하는 테스트 세트에서 79.08%의 정확도를 달성했습니다. 번역체 분류기 실험에 대한 상세한 결과와 제거 연구는 부록 H.2에 제시되어 있습니다. 이 분류기를 통해 번역체와 자연 발생 텍스트를 구분함으로써 LLM의 생성 품질을 평가할 수 있으며, 이는 모델이 진정한 언어 출력을 생성하는 성능에 대한 통찰력을 제공합니다.

[10] https://huggingface.co/microsoft/mdeberta-v3-base

LLM의 생성 품질 평가 우리는 Sea-Bench(Nguyen et al., 2023)의 자연어, 일반적, 안전 관련 질문에 대한 답변을 생성함으로써 9개 동남아시아 언어에서 LLM의 생성 품질을 평가합니다. 표 2a에서 볼 수 있듯이, 광범위한 언어 커버리지를 가지지만 동남아시아 언어에 덜 집중하는 대규모 언어 모델(예: AYA-101(Üstün et al., 2024), GPT-4(OpenAI et al., 2024), mT0(Muennighoff et al., 2023; Xue et al., 2021), Llama3 (AI@Meta, 2024) 등이 해당됩니다. 반면, 언어 커버리지는 좁지만 SEA 언어에 더 집중한 모델들, 예를 들어 Cendol-Llama2 (Cahyawijaya et al., 2024b), Sailor (Dou et al., 2024), AYA-23 (Aryabumi et al., 2024), SEA-LION (Singapore, 2023)과 같이 SEA 언어에 더 집중하는 모델들은 35% 이상의 시간 동안 자연스러운 문장을 생성합니다.

그러나 SEA 언어에서 가장 적은 번역체 생성을 보이는 LLM인 SEA-LION조차도 자연스러운 SEA 문장을 57.71%의 확률로만 생성하여, SEA 언어에서 자연스러운 문장을 생성하는 데 있어 상당한 품질 격차가 있음을 보여준다. 표 2b에서 볼 수 있듯이, 번역어 문제는 동남아시아 언어마다 다릅니다. 타갈로그어(TGL), 버마어(MYA), 말레이어(ZLM)와 같은 언어는 번역어 문제가 더 심각하여, 기존 LLM이 자연스러운 문장을 생성하는 비율은 각각 11.58%, 19.47%, 22.24%에 불과합니다. 이는 동남아시아 언어의 언어적 다양성과 복잡성을 보다 효과적으로 해결하기 위해 LLM을 더욱 개선할 필요가 있음을 강조한다.

5 토론

5.1 동남아시아의 자원 격차

SEACrowd는 동남아시아(SEA)에서 사용되는 1,308개 언어 중 980개 언어(74.9%)를 커버합니다. 이러한 높은 커버리지에도 불구하고, SEACrowd의 언어별 데이터 분포는 매우 긴 꼬리 분포를 보이며, 700개 이상의 언어가 단 1개 또는 2개의 데이터셋만을 보유하고 있습니다.

그리고 20개 이상의 데이터셋을 보유한 언어는 단 23개에 불과합니다. 이러한 상대적으로 덜 대표되는 언어들은 일반적으로 어휘집 형태(Asgari et al., 2020; List et al., 2022)나 라벨링되지 않은 데이터 형태(Leong et al., 2022; Kudugunta et al., 2024; Nguyen et al., 2024)로만 존재합니다. SEACrowd의 기존 작업들은 여전히 소수 언어만을 다루고 있습니다. 예를 들어, 감정 분석 데이터는 22개 언어에서만 이용 가능하며, 명명체 인식(NER) 데이터는 고작 17개 언어에서만 제공됩니다. 더욱이 텍스트를 넘어선 모달리티에 대해서는 SEA 자원이 극히 부족합니다. SEA 원주민 언어의 약 90%는 음성 및 시각-언어(VL) 데이터셋 모두를 갖추지 못하고 있습니다.

품질 SEACrowd 데이터셋의 78.7%는 동료 평가를 거친 학술지에 게재되었으며, 대부분의 데이터는 외부 검증을 거쳤습니다. SEACrowd 데이터셋의 전반적인 품질은 그림 5b에 나타내었습니다. 저자들이 보고한 데이터 구축 방법을 데이터 수집 방법(즉, 데이터 소스)과 레이블 주석 검증(즉, 품질 관리)을 모두 고려하여 정리하였습니다. SEACrowd 데이터셋의 약 19%는 기계 생성 및 기계 번역 주석을 사용한 반면, 80% 이상은 온라인 텍스트(예: 웹 크롤링) 및 전문가 생성 방식으로 획득되었습니다. 라벨 주석 검증 측면에서는 데이터셋의 62.4%가 완전 수동 검증을 거쳤으며, 나머지 부분은 부분 검증 및 자동 검증을 거쳤습니다. 이러한 통계는 표면적인 데이터셋 수집 품질에 대한 초기 지표만을 제공하며 정확한 품질을 반드시 반영하지는 않습니다. SEACrowd 내 상세 품질 지표(예: 주석자 간 일치도 점수)를 보고한 데이터셋은 극소수(6%)에 불과합니다. 향후 연구를 위한 심층적인 조사가 필요합니다.

문화적 관련성 동남아시아 데이터의 자원 격차는 문화적 측면까지 확대되며, 잘못된 표현은 문화적 전유나 고정관념과 같은 불쾌한 행동으로 이어질 수 있다(Evans et al., 2020; Glotov, 2023). SEA 데이터 세트의 문화적 관련성을 평가하기 위한 대리 지표로, SEACrowd 평가에 사용된 259개 데이터 하위 집합을 데이터 출처 기준으로 수동 선별했습니다. 구체적으로 1) 타 언어에서 번역된 자료, 2) 현지 출처에서 크롤링된 자료, 3) 문화적 관련성 확보를 위해 수작업으로 제작된 자료로 분류했습니다. 그림 5c에서 약 70%는 문화적 관련성이 부족한 것으로 나타났는데, 이는 대부분 영어 출처에서 기계 번역된 자료이기 때문이다. 약 20%는 현지 출처에서 가져온 자료이며
뉴스, 소셜 미디어 또는 기타 지역 매체에 대한 접근성이 여전히 기대치에 미치지 못하는 것으로 나타났습니다.
문화적으로 관련성이 있는 데이터를 포함할 가능성이 있습니다.

나머지 10%만이 문화적 관련성을 고려하도록 설계되었으며, 이는 소수 언어에 대한 LLM의 심각한 문화 이해 부족을 강조한 연구들(Kabra et al., 2023; Koto et al., 2023a; Wibowo et al., 2023; Liu et al., 2024; Koto et al., 2024)에서 파생되었습니다.

5.2 결론 및 향후 연구 방향

동남아시아는 매우 다양한 언어와 문화를 지닌 지역으로, 대다수 주민이 영어를 주 언어로 사용하지 않습니다. 영어 중심 AI의 유용성은 동남아시아 사용자의 대다수에게 제한적이며, 특히 의료 및 교육과 같은 핵심 분야에서 더욱 그렇습니다. SEACrowd를 통해 우리는 동남아시아의 AI 현황을 탐색하고 동남아시아 언어 기반 AI 모델의 자원, 평가, 자연스러움 분석 분야의 격차를 해소했습니다. 또한 본 프로젝트는 오픈소스 연구 커뮤니티를 육성하여 데이터시트와 데이터 로더의 추가 및 유지 관리를 지속하고 동남아시아의 AI 연구 개발을 주도할 것입니다.

그럼에도 동남아시아의 AI 개발은 다양한 이해관계자들의 집중적인 노력이 필요하며, 이들은 1,300개 이상의 지역 언어를 AI 모델에 통합하는 데 있어 서로 다른 우선순위를 가질 수 있습니다. 향후 동남아시아의 AI 개발은 두 가지 핵심 지표를 우선시해야 한다는 것이 우리의 연구 결과입니다: 1) 잠재적 유용성 및 2) 자원 형평성.[11]

잠재적 효용 잠재적 효용은 현재 효용과 이상적 효용 간의 차이로 정의되며, 여기서 모델 능력은 효용의 대리 변수로 작용한다. 잠재적 효용을 기준으로 볼 때, 당연하게도 싱가포르에서 사용되는 영어와 중국어를 제외한 국가 언어, 즉 인도네시아어(IND), 버마어(MYA), 베트남어(VIE), 태국어(THA), 필리핀어(FIL), 크메르어(KHM), 말레이어(ZLM), 라오어(LAO)의 발전이 가장 큰 혜택을 가져올 것이다. 이 중 말레이어, 버마어, 필리핀어 AI 생성 출력의 자연스러움에서 두드러진 격차를 확인했다(§4.2). 해당 언어에 대한 자원 구축에 집중된 노력이 효용 측면에서 가장 큰 변화를 가져올 수 있다. 국가 언어 외에도 자바어(JAV), 순다어(SUN), 몽족어(HMN) 등 대규모 사용자 기반을 가진 성장 중인 지역 언어 또는 방언을 키우는 것이 핵심이다.

자원 형평성 자원 형평성은 기존 자원 가용성과 이상적인 자원 가용성 간의 격차로 정의됩니다(그림 6). 많은 지역 언어 또는 방언이 여전히 기대되는 수준에 미치지 못하는 것으로 나타났습니다.

[11] https://github.com/SEACrowd/globalutility

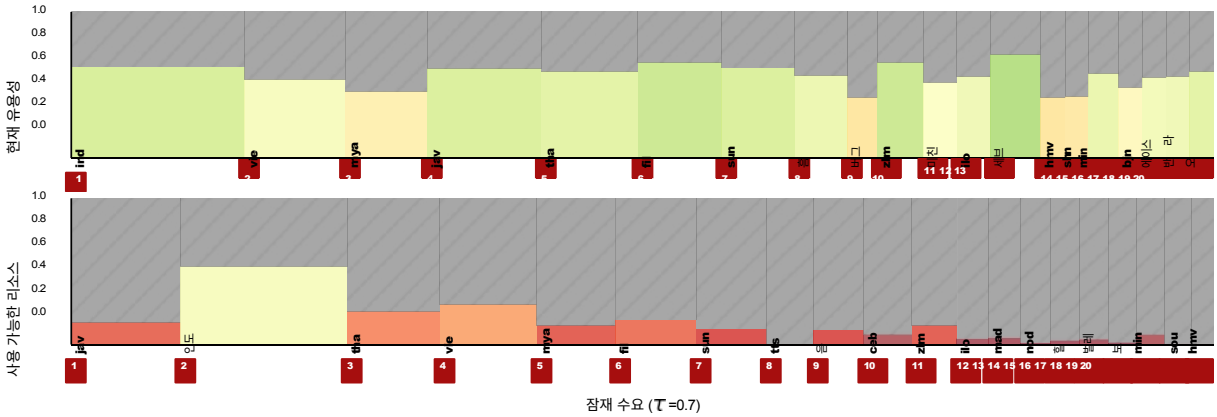


그림 6: **(위)** 현재 유용성과 **(아래)** 자원 가용성을 기반으로 한 SEA 언어 우선순위화. 언어는 누락된 잠재적 영역 크기의 내림차순으로 순위가 짐 .

. 여기에는 북동부 태국어(TTS), 북부 태국어(NOD), 몽도어(HMV), 남부 태국어(SOU), 세부아노어(CEB), 일로카노어(ILO) 등이 포함됩니다. 이러한 격차를 줄이기 위한 노력은 해당 언어의 보존뿐만 아니라 해당 언어 사용자들의 문화유산이 지속되도록 보장할 것이다. 수요에 대한 다양한 가중치를 적용한 동남아시아 언어 우선순위 지정의 자세한 내용은 부록 I에서 확인할 수 있다.

이러한 지표를 개선하기 위해 해당 지역의 정부 및 산업 리더들은 국가 언어와 지역 방언 모두에 대한 지역 언어 역량을 강화하기 위한 연구 개발 활동에 투자해야 합니다. 이를 위해서는 공개 데이터 수집에 대한 자금 지원과 지역 사회와의 협력을 통해 지역 언어의 자원 격차를 해소하는 방안이 포함될 수 있습니다. 또한 포용적 AI 모델을 기반으로 수익성 있는 활용 사례를 촉진하고, 데이터 작업자에 대한 공정하고 책임감 있는 보상 체계를 장려하며, 데이터 소유자, AI 및 애플리케이션 개발자 간의 상생 협력 모델을 구축하는 등 장기적이고 지속 가능한 전략이 필요합니다.

감사의 말

저희는 놀라운 기여자분들께 감사드립니다: 조슈아 스퍼겔, 티젱 위, 파린 타팟 펭푼, 이샨 진달, 무함마드 사트리오, 지펑 장, 바비쉬 파와, 하료 아크바리안토 위보워, 히로키 노모토, 요하네스 시기트 푸르노모 W.P., 아흐마드 파탄 히다야툴라, 브라이언 윌리, 루히야 파라디시 위디아푸트리, 라피프 라바니, 파와즈 마이다, 마노즈 카트리, 수프리아디 수프리아디, 비라치 손렐트람바니치, 파바리스 루앙추티포판, 에를란드 힐만 푸아디, 메가 프란시스카, 리차디 사판, 그리고 카밀라 조닌 코스메에게 SEACrowd를 위한 데이터 시트 제출과 데이터 로더 구현에 대한 노고에 감사드립니다.

본 연구는 싱가포르 국가연구재단(NRF)의 AI 싱가포르 프로그램, 홍콩과학기술대학교(HKUST) 박사 과정 장학금, 그리고 홍콩 연구기금회(RGC)의 PF20-43679 홍콩 박사 과정 장학금 제도의 지원을 받았습니다. JMI는 필리핀 국립대학교와 영국 바스 대학교의 UKRI 책임감 있고 투명한 AI 박사 과정 교육 센터[EP/S023437/1]의 지원을 받았습니다. 또한, 상용 기반 모델인 Command-R을 활용한 실험 수행을 가능하게 해준 연구 지원금을 제공해 주신 Cohere For AI에 깊은 감사를 표합니다.

제한 사항

우리의 연구는 거의 1,000개의 동남아시아 언어를 다루지만, 모어에 속하는 것으로 간주되는 많은 방언들이 평가 벤치마크에서 누락되어 있습니다. 예를 들어 말레이어의 경우 표준 말레이어(ZSM)만 평가되며, 사라왁 말레이어(ZLM-SAR)와 같은 다른 방언들은 평가 대상이 아닙니다. 또한, 대부분의 데이터셋에는 SEA 언어 사용에서 흔히 나타나는 언어 현상인 코드 스위칭 텍스트가 포함되어 있지 않습니다(Aji et al., 2023). 게다가, 다양한 평가 작업의 언어 커버리지는 상당히 차이가 납니다. 예를 들어, NLP 작업은 총 34개 언어를 다루는 반면, VL 작업은 단 4개 언어만 다룹니다. 이러한 한계를 해결하는 것은 동남아시아를 더 잘 대표하는 데 필수적이며, 향후 연구에서는 이러한 측면을 우선적으로 고려할 것을 강력히 권장합니다.

윤리 선언문

SEA 언어 평가 벤치마크 개발 과정에서 윤리적 고려 사항이 포괄적으로 반영되도록 여러 단계를 거쳤습니다. 첫째, 본 벤치마크에 사용된 데이터는 공개적으로 이용 가능한 자료에서 수집되었으며, 이는 데이터 개인정보 보호와 관련된 법적·윤리적 기준을 준수함을 의미합니다. 해당되는 경우 데이터 제공자로부터 명시적 동의를 얻었습니다. 또한 본 벤치마크에 활용된 모든 데이터셋과 자료는 각자의 라이선스에 따라 사용되었습니다. 둘째, 본 벤치마크는 포괄성을 지향하여 현재 언어학 자료에서 상대적으로 덜 다루어지는 언어를 포함해 동남아시아의 다양한 언어를 대표하도록 구성되었습니다. 마지막으로, 데이터 수집, 벤치마크 개발, 평가 방법론 등 연구 과정 전체가 완전히 오픈소스로 공개되었으며, 재현성과 책임성을 보장하기 위해 투명하게 문서화되었습니다.

참고문헌

데이비드 아델라니, 제수조바 알라비, 안젤라 팬, 줄리아 크로이처, 샤오유 쉔, 마첼 리드, 다나 루이터, 디트리히 클라코우, 피터 나벤데, 어니 창, 타주딘 그와다베, 프레시아 사키, 보나벤투르 F. P. 도수, 크리스 에메주, 콜린 렁, 마이클 부크만, 샴수딘 무함마드, 구요 자르소, 오린 유수프, 안드레 니용가보 루붕고, 질 하셰메, 에릭 피터 와이라갈라, 무함마드 우마이어 나시르, 벤자민 아지바데, 툰데 아자이, 이본 기타우, 제이드 애벗, 모하메드 아메드, 밀리센트 오치엔, 안울루와포 아레무, 페레즈 오가요, 조너선 무키이비, 파투마타 우오바 카보레, 갓슨 칼리페, 데르게네 음바예, 알라세라 오귀스트 타포, 빅투아르 멤조캄 코아그네, 에드윈 문코-부아벵, 발렌시아 바그너, 이드리스 압둘무민, 아요델레 아워코야, 해피 부자아바, 블레싱 시반다, 안디스와 부쿨라, 샘 만탈루. 2022a. A few thousand trans-lations go a long way! leveraging pre-trained mod-els for African news translation. In *Proceedings of the 2022 Conference of the North American Chap-ter of the Association for Computational Linguistics: Human Language Technologies,* pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassi-lyev, Jesujoba Alabi, Yanke Mao, Haonan Gao, and En-Shiun Lee. 2024. SIB-200: 200개 이상의 언어 및 방언에서 주제 분류를 위한 간단하고 포괄적이며 대규모 평가 데이터셋. *제18회 유럽계산언어학회(ACL) 학술대회 논문집(제1권: 장문 논문),* 226–245쪽, 몰타 세인트줄리안스. 유럽계산언어학회(ACL).

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-

미셸, 콘스탄틴 리뇨스, 제수조바 알라비, 샴수딘 무함마드, 피터 나벤데, 셰이크

M. 밤바 디온, 안디스와 부쿨라, 루웨이터 마부야, 보나벤투르 F. P. 도수, 블레싱 시반다, 해피 부자아바, 조너선 무키이비, 갓슨 칼리페, 데르게네 음바이, 아멜리아 테일러, 파투마타 카보레, 크리스 치넨예 에메주, 아누올루와포 아레무, 페레즈 오가요, 캐서린 기타우, 에드윈 문코-부아벵, 빅토르 멤조캄 코아그네, 알라세라 오귀스트 타포, 테보고 마쿠콰, 부코시 마리바테, 음본잉 치아제 엘비스, 타주딘 그와다베, 토신 아데우미, 오레바오게네 아히아, 조이스 나카툼바-나벤데. 2022b. MasakhaNER 2.0: 명명된 엔티티 인식을 위한 아프리카 중심 전이 학습. *2022년 자연어 처리 경험적 방법 컨퍼런스 논문집,* 4488–4508쪽, 아부다비, 아랍에미리트. 계산언어학회.

데이비드 이페올루와 아델라니, 제이드 애벗, 그레이엄 노이빅, 다니엘 드수자, 줄리아 크로이처, 콘스탄틴 리그노스, 체스터 팔렌미셸, 해피 부자바, 슈루티 리즈와니, 세바스찬 루더, 스티븐 메이휴, 이스라엘 아베베 아지메, 샴수딘 H. 무함마드, 크리스 치넨예 에메주에, 조이스 나카툼바-나벤데, 페레즈 오가요, 아레무 아누올루와포, 캐서린 기타우, 데르게네 음바예, 제수조바 알라비, 세이드 무히 예맘, 타주딘 라비우 그와다베, 이그나티우스 에제아니, 루붕고 안드레 니용가보, 조너선 무키이비, 베라 오티엔데, 이로로 오리페, 데이비스 데이비드, 삼바 응옴, 토신 아데우미, 폴 레이슨, 모페톨루와 아데예미, 제럴드 무리우키, 엠마누엘 아네비, 치아마카 추쿤네케, 은키루카 오두, 에릭 피터 와이라갈라, 사무엘 오예린데, 클레멘시아 시로, 토비우스 사울 바테사, 테밀롤라 올로예데, 이본 웜부이, 빅터 아킨-오데, 데보라 나바게레카, 모리스 카투시임, 아요-델레 아워코야, 무하마다네 음부프, 디보라 게브레요한네스, 헤녹 틸라예, 켈레치 은와이케, 데-가가 월데, 압둘라예 파예, 블레싱 시반다, 오레-바오게네 아히아, 보나벤투르 F. P. 도수, 켈레치 오게지, 티에르노 이브라히마 디오프, 압둘라예 디알로, 아데왈레 아킨파데린, 텐다이 마렌게레케, 살로메이 오세이. 2021. MasakhaNER: 아프리카 언어들을 위한 명명체 인식. *계산언어학회지,* 9:1116–1131.

데이비드 이페올루와 아델라니, 마렉 마시아크, 이스라엘 아베베 아지메, 제수조바 알라비, 아트나푸 람베보 톤자, 크리스틴 므와세, 오두나요 오군데포, 보나벤투르

F. P. 도수, 아킨툰데 올라디포, 도린 닉스도르프, 크리스 치넨예 에메주, 사나 알-아자위, 블레싱 시반다, 데이비스 데이비드, 롤웨투 은돌레라, 조너선 무키이비, 툰데 아자이, 타티아나 모테우, 브라이언 오디-암보, 아브라함 오워둔니, 나에메카 오비에푸나, 무히딘 모하메드, 샴수딘 하산 무함마드, 테쇼메 물루게타 아바부, 사히드 압둘라히 살라후딘, 메사이 게메다이게주, 타주딘 그와다베, 이드리스 압둘무민, 마흘레트 타예, 올루와부사요 아워요미, 이야누올루와 쇼데, 톨루로페 아델라니, 하비바 압둘가니유, 압둘하킴 오모타요, 아데톨라 아데코, 아비브 아폴라비, 안울루와포 아레무, 올라렐와주 사무엘, 클레멘시아 시로, 왕가리 키모토, 오니에카치 오그부, 치네두

음보누, 치아마카 추쿠네케, 사무엘 파니조, 제시카 오조, 오잉칸솔라 아워산, 타데세 케베데, 토아둠 사리 사카요, 파멜라 냐치네, 프리드모어 시두메, 오린 유수프, 마르디야 오두볼레, 칸다 치누, 우센 키마누카, 티나 디코, 시얀다 은자카마, 시노도스 니구세, 압둘메지드 조하르, 샤피 모하메드, 푸아드 미레 하산, 모게스 아메드 메하메드, 에브라르 응가비레, 쥴스 쥴스, 이반 센쿵구, 폰투스 스테네토프. 2023. Masa k haNEWS: 아프리카 언어들을 위한 뉴스 주제 분류. *제13회 국제 자연어 처리 공동 학술대회 및 제3회 아시아-태평양 계산언어학회 학술대회 논문집 (제1권: 장문 논문)*, 144–159쪽, 누사 두아, 발리. 계산언어학회.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Ayu Purwarianti. 2023. 모듈형 다국어 언어 모델의 모호한 한계. *ICLR Tiny Papers 2023*.

무함마드 파리드 아딜라주아르다, 사그닉 무케르지, 프라디움나 라바니아, 시단트 싱, 아슈토시 드위베디, 알함 피크리 아지, 재키 오닐, 아슈토시 모디, 모노짓 차우드리. 2024. LLM에서 "문화" 측정 및 모델링을 향하여: 서베이. *프리프린트*, arXiv:2403.15412.

AI@Meta. 2024. Llama 3 모델 카드.

Alham Fikri Aji, Jessica Zosa Forde, Alyssa Marie Loo, Lintang Sutawika, Skyler Wang, Genta Indra Winata, Zheng-Xin Yong, Ruochen Zhang, A. Seza Dog˘ruöz, Yin Lin Tan, Jan Christian Blaise Cruz. 2023. 다국어 사용과 언어 다양성에서 얻은 통찰력을 통한 동남아시아의 NLP 현황. *제13회 국제 자연어 처리 공동 학회 및 제3회 아시아 태평양 컴퓨터 언어학 협회 학회: 튜토리얼 초록*, 8–13쪽, 누사 두아, 발리. 컴퓨터 언어학 협회.

알함 피크리 아지, 겐타 인드라 위나타, 파즈리 코토, 사무엘 카야위자야, 아데로마도니, 라흐마드 마헨드라, 케말 쿠르니아완, 데이비드 뫼르자디, 라디티오 에코 프라소조, 티모시 볼드윈, 제이 한 라우, 세바스찬 루더. 2022. 한 나라, 700개 이상의 언어: 인도네시아의 소수 언어 및 방언을 위한 NLP 과제. *제60회 전산언어학회 연차 총회 논문집 (제1권: 장문 논문)*, 7226–7249쪽, 더블린, 아일랜드. 전산언어학회.

바드르 알카미시, 무함마드 엘노크라시, 마이 알카미시, 모나 디아브. 2024. 대규모 언어 모델의 문화적 정렬성 조사. *사전 출판물*, arXiv:2402.13231.

에브테삼 알마즈루에이, 함자 알오베이들리, 압둘아지즈 알샴시, 알레산드로 카펠리, 룩산드라 코조카루, 메루안 데바, 에티엔 고피네, 다니엘 헤슬로, 쥘리앙 로네, 콴탱 말라르티크, 바드레딘 Noune, Baptiste Pannier, Guilherme Penedo. 2023. Falcon-40B: 최첨단 성능을 자랑하는 오픈 소스 대규모 언어 모델.

로사나 아르딜라, 메건 브랜슨, 켈리 데이비스, 마이클 콜러, 조쉬 마이어, 마이클 헨레티, 루벤 모라이스, 린지 손더스, 프랜시스 타이어스, 그리고르 웨버. 2020. 공통 음성: 대규모 다국어 음성 코퍼스. *제12회 언어 자원 및 평가 컨퍼런스 논문집*, 4218–4222쪽, 프랑스 마르세유. 유럽 언어 자원 협회.

미켈 아르테체, 세바스티안 루더, 다니 요가타마. 2020. 단일 언어 표현의 다언어 간 이전 가능성에 관하여. *제58회 전산언어학회 연차 총회 논문집*, 4623–4637쪽, 온라인. 전산언어학회.

비라트 아리아부미, 존 당, 드와라크 탈루푸루, 사우라브 대시, 데이비드 카이루즈, 헝위 린, 바라트 벤키테시, 매들린 스미스, 켈리 마르키시오, 세바스티안 루더, 아시르 로카텔리, 줄리아 크로이처, 닉 프로스트, 필 블런섬, 마르지 파다에, 아흐메트 우스툰, 사라 후커. 2024. Aya 23: 다국어 발전을 위한 오픈 웨이트 재공개. *사전 인쇄본*, arXiv:2405.15032.

아사 이 아카리, 스네하 쿠두군타, 신얀 벨로시티 유, 테라 블레빈스, 힐라 B 고넨, 마첼 리드, 율리아 츠베트코프, 세바스찬 루더, 한나네 하지시르지. 2023. BUFFET: 다국어 소량 데이터 전이를 위한 대규모 언어 모델 벤치마킹. *사전 인쇄본*, arXiv:2305.14857.

에사네딘 아스가리, 파비엔 브라운, 벤자민 로스, 크리스토프 링글스테터, 모하마드 모프라드. 2020. UniSent: 1000개 이상의 언어를 위한 보편적 적응형 감정 어휘집. *제12회 언어 자원 및 평가 컨퍼런스 논문집*, 4113–4120쪽, 프랑스 마르세유. 유럽 언어 자원 협회.

락스미타 위디아 아스투티, 유니타 사리, 수프라토. 2023. 트위터 소셜 미디어 데이터를 위한 트랜스포머 기반 코드 혼합 감정 분석. *국제 고급 컴퓨터 과학 및 응용 저널*, 14(10).

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, Michael Auli. 2021. Xls-r: 대규모 자체 지도 학습 기반 다국어 음성 표현 학습. *사전 인쇄본*, arXiv:2111.09296.

루카스 반다르카르, 데이비스 리앙, 벤자민 뮐러, 미켈 아르테체, 사티아 나라얀 슈클라, 도널드 후사, 나만 고얄, 아비난단 크리슈난, 루크 제틀모이어, 마디안 카브사. 2023. 벨레벨레 벤치마크: 122개 언어 변형의 병렬 독해 데이터셋. *arXiv 사전 인쇄본 arXiv:2308.16884*.

Damian Blasi, Antonios Anastasopoulos, Graham Neubig. 2022. 세계 언어 간 언어 기술 성능의 체계적 불평등. *제60회 전산언어학회 연례 회의 논문집 (제1권: 장문 논문)*, 5486–5505쪽, 아일랜드 더블린. 전산언어학회.

Andrea Burns, Donghyun Kim, Derry Wijaya, Kate Saenko, Bryan A Plummer. 2020. 다국어 표현을 확장하여 학습하는 비전-언어 작업용 다중 언어 표현 학습( ). *컴퓨터 비전–ECCV 2020: 제16회 유럽 컨퍼런스, 영국 글래스고, 2020년 8월 23–28일, 논문집, 파트 IV 16*, 197–213쪽. Springer.

사무엘 카야위자야, 알함 피크리 아지, 홀리 로베니아, 겐타 인드라 위나타, 브라이언 윌리, 라흐마드 마헨드라, 파즈리 코토, 데이비드 모엘자디, 카리사 빈센티오, 아데 로마도니, 아유 푸르와리안티. 2022. 누사크라우드: 인도네시아어에서 개방적이고 재현 가능한 자연어 처리 연구를 위한 제안. *사전 인쇄본*, arXiv:2207.10524.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, Pascale Fung. 2024a. 대규모 언어 모델에서의 고차원 인간 가치 표현. *arXiv 사전 인쇄본 arXiv:2404.07900*.

사무엘 카야위자야, 홀리 로베니아, 알함 피크리 아지, 겐타 위나타, 브라이언 윌리, 파즈리 코토, 라흐마드 마헨드라, 크리스티안 위비소노, 아데 로마도니, 카리사 빈센티오, 제니퍼 산토소, 데이비드 모엘-자디, 카야 위라완, 프레데리쿠스 후디, 무함마드 사트리오 위카크소노, 이반 파르모낭간, 이카 알피나, 일함 피르다우스이 푸트라, 삼술 라흐마다니, 유-리안티 오낭, 알리 세프티안드리, 제임스 자야, 카우스투브 돌레, 아리 수르야니, 리프키 아피나 푸트리, 단 수, 키스 스티븐스, 마데 닌디야타마 니티아샤, 무함마드 아딜라주아르다, 라이언 하디위자야, 랴안디토 디안다루, 티젱 위, 비토 기파리, 웬량 다이, 옌 쉬, 디야 다마푸스피타, 하리오 위보워, 쿡 토, 이치-와눌 카로 카로, 티라나 파티아노사, 지웨이 지, 그레이엄 노이빅, 티모시 볼드윈, 세바스찬 루더, 파스칼 펑, 헤리 수자이니, 사크리아니 사크티, 아유 푸르와리안티. 2023a. NusaCrowd: 인도네시아어 자연어 처리 자원을 위한 오픈 소스 이니셔티브. *계산언어학회 학술대회 논문집: ACL 2023*, 13745–13818쪽, 캐나다 토론토. 계산언어학회.

사무엘 카야위자야, 홀리 로베니아, 파즈리 코토, 데아 아디스타, 엠마누엘 데이브, 사라 옥타비안티, 살사빌 아크바르, 존슨 리, 누르 샤디에크, 쳉 와완 쳉고로, 하눙 리누위, 브라이언 윌리, 갈리흐 무리단, 겐타 위나타, 데이비드 모엘자디, 알함 피크리 아지, 아유 푸르와리안티, 파스칼 펑. 2023b. 누사라이트스: 저조한 대표성과 극히 부족한 자원을 가진 언어를 위한 고품질 코퍼스 구축. *제13회 국제 자연어 처리 공동 학술대회 및 제3회 아시아 태평양계산언어학회 학술대회 논문집(제1권: 장문 논문)*, 921–945쪽, 누사 두아, 발리. 컴퓨터 언어학회.

사무엘 카야위자야, 홀리 로베니아, 파즈리 코토, 리프키 아피나 푸트리, 엠마누엘 데이브, 존슨 리, 누르 샤디에크, 와완 첸고로, 살사빌 마울라나 아크바르, 무함마드 이자 마헨드라, 데아 아니사얀티 푸트리, 브라이언 윌리, 겐타 인드라 위나타, 알함 피크리 아지, 아유 푸르와리안티, 파스칼 펑. 2024b. 센돌: 인도네시아어에 대한 공개 교육 조정 생성형 대규모 언어 모델. *사전 인쇄본*, arXiv:2404.06138.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kun-coro, Sebastian Ruder, Zhi Yuan Lim, Syafri Ba-har, Masayu Khodra, Ayu Purwarianti, Pascale Fung. 2021. IndoNLG: 인도네시아어 자연어 생성을 평가하기 위한 벤치마크 및 자원. *2021년 자연어 처리의 경험적 방법 컨퍼런스 논문집*, 8875–8898쪽, 온라인 및 도미니카 공화국 푼타 카나. 컴퓨터 언어학 협회.

재스퍼 카일 카타팡과 모세스 비스페라스. 2023. 타갈로그어와 영어 시나리오에서의 감정 기반 도덕성(EMoTES-3K): 행동의 (비)도덕성을 설명하기 위한 병렬 코퍼스. *제3회 디지털 인문학을 위한 자연어 처리 국제학술대회 및 제8회 우랄어군을 위한 계산언어학 국제워크숍 공동 학술대회 논문집*, 1–6쪽, 일본 도쿄. 계산언어학회.

원활한 소통, 로익 바로, 정유안, 마리아노 코라 메글리오리, 데이비드 데일, 동닝, 폴-앙브루아즈 뒤켄, 하디 엘사하르, 공홍유, 케빈 헤퍼넌, 존 호프만, 크리스토퍼 클라이버, 리펑웨이, 다니엘 리히트, 장 마이야르, 앨리스 라코토아리송, 카우식 람 사다고판, 기욤 웬제크, 이단 예, 바피 아쿨라, 펑젠 천, 나지 엘 하셈, 브라이언 엘리스, 가브리엘 메히아 곤잘레스, 저스틴 하하임, 프랑팁 한산티, 러스 하우스, 버니 황, 황민재, 이나가마 히로후미, 소미야 자인, 엘라헤 칼바시, 아만다 칼렛, 일리아 쿨리코프, 재니스 램, 다니엘리, 마쉬타이, 루슬란 마블류토프, 벤자민 펠로켄, 모하메드 라마단, 아비네쉬 라마크리슈난, 안나 선, 케빈 트란, 투안 트란, 이고르 투파노프, 비쉬 보게티, 칼리 우드, 양이린, 유보카이, 피에르 앤드루스, 칸 발리오글루, 마르타 R. 코스타-후사, 오누르 셀레비, 마하 엘바야드, 신시아 가오, 프란시스코 구즈만, 저스틴 카오, 앤 리, 알렉산드르 무라치코, 후안 피노, 스라비아 포푸리, 크리스토프 로퍼스, 사피야 살림, 홀거 슈웽크, 페이든 토마셀로, 왕창한, 제프 왕, 스카일러 왕. 2023. Seamlessm4t: 대규모 다국어 및 다중 모달 기계 번역. *사전 인쇄본*, arXiv:2308.11596.

알렉시스 코노, 알렉세이 바예프스키, 로난 콜로베르, 압델라흐만 모하메드, 마이클 아울리. 2021. 음성 인식을 위한 비지도 교차 언어 표현 학습. *Interspeech 2021 논문집*, 2426–2430쪽.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, Daan van Esch, Vera Axelrod, Simran Khanuja, Jonathan Clark, Orhan Firat, Michael Auli, Sebastian Ruder, Jason Riesa, Melvin Johnson. 2022. X TREME-S: 평가를 통한 다국어 음성 표현. *Proc. Interspeech 2022,* 3248–3252쪽.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, Veselin Stoyanov. 2018. XNLI: 교차 언어 문장 표현 평가. *2018년 자연어 처리 경험적 방법 컨퍼런스 논문집,* 2475–2485쪽, 벨기에 브뤼셀. 컴퓨터 언어학 협회.

마르타 R. 코스타-후사, 제임스 크로스, 오누르 첼레비, 마하 엘바야드, 케네스 히필드, 케빈 헤퍼넌, 엘라헤 칼바시, 재니스 램, 다니엘 리히트, 장 마이야르, 안나 선, 스카일러 왕, 기욤 벤제크, 알 영블러드, 바피 아쿨라, 로익 바를로, 가브리엘 메히아 곤살레스, 프랑팁 한산티, 존 호프만, 세마리 재럿, 카우식 람 사다고판, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Jeff Wang, and N. L. L. B. 팀. 2024. 200개 언어로 확장된 신경망 기계 번역. *네이처.*

Raj Dabre, Himani Shrotriya, Anoop Kunchukuttan, Ratish Puduppully, Mitesh Khapra, Pratyush Kumar. 2022. IndicBART: 인도계 언어 자연어 생성을 위한 사전 훈련 모델. In *Findings of the Association for Computational Linguistics: ACL 2022,* pp. 1849–1863, Dublin, Ireland. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, Thien Huu Nguyen. 2023. Okapi: 인간 피드백을 통한 강화 학습으로 다중 언어에 적용된 지시어 튜닝 대형 언어 모델. *arXiv 전자 인쇄본*, arXiv–2307.

대원량, 리준난, 리동쉬, 안토니 멩화퉁, 자오준치, 왕웨이성, 리보양, 펑파스칼 엔, 호이스티븐. 2024. Instructblip: 지시문 튜닝을 통한 범용 시각-언어 모델 구축. *신경정보처리시스템 발전*, 36.

Robert Dorfman. 1979. 지니 계수를 위한 공식. *경제학 및 통계 리뷰*, 146–149쪽.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Ji-ahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: 동남아시아를 위한 오픈 언어 모델. *사전 인쇄본*, arXiv:2404.03608.

Matthew S. Dryer 및 Martin Haspelmath, 편집자. 2013. *WALS Online (v2020.3)*. Zenodo.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph 외. 2023. 언어 모델에서 주관적인 글로벌 의견의 표현을 측정하기 위해. *arXiv 사전 인쇄본 arXiv:2306.16388*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World. Twenty-fourth edition.* Dallas, Texas: SIL Interna-tional.

압틴 에브라히미, 마누엘 마거, 아르투로 온세바이, 비슈라브 차우다리, 루이스 치루조, 안젤라 팬, 존 오르테가, 리카르도 라모스, 아네트 리오스, 이반 블라디미르 메사 루이스, 구스타보 기메네스-루고, 엘리자베스 마거, 그레이엄 노이빅, 알렉시스 팔머, 롤란도 코토-솔라노, 탕 부, 카타리나 칸. 2022. AmericasNLI: 진정한 저자원 언어에서 사전 훈련된 다국어 모델의 제로샷 자연어 이해 평가. *제60회 컴퓨터 언어학회 연차 총회 논문집(제I 권: 장문 논문),* 6279–6299쪽, 아일랜드 더블린. 컴퓨터 언어학회.

Alexander Elias. 2018. Lio and the central flores lan-guages. *Leiden: Leiden University Master thesis*.

리안 에반스, 크리스타사니 터너, 켈리 앨런. 2020. "좋은 교사"와 "좋은 의도": 문화적 반응형 교육법의 오용. *도시 학습, 교수 및 연구 저널*, 15(1):51–73.

크 리스천 페더만, 톰 코크미, 잉 신. 2022. NTREX-128 – 128개 언어 기계번역 평가를 위한 뉴스 테스트 참조 자료. *제I회 다국어 평가 확장 워크숍 논문집*, 21–24쪽, 온라인. 컴퓨터 언어학회.

팀닛 게브루, 제이미 모건스턴, 브리아나 베키오네, 제니퍼 워트먼 본, 한나 왈락, 할 도메 3세, 케이트 크로포드. 2021. 데이터셋을 위한 데이터시트. *ACM 커뮤니케이션즈*, 64(12):86–92.

Gregor Geigle, Abhay Jain, Radu Timofte, Goran Glavaš. 2023. mblip: 효율적인 다국어 비전-LLM 부트스트래핑. *arXiv*, abs/2307.06930.

제마 팀, 토마스 메스나르, 캐시디 하딘, 로버트 다다시, 수르야 부파티라주, 슈레야 파탁, 로랑 시프레, 모르간 리비에르, 미히르 산자이 칼레, 줄리엣 러브, 푸야 타프티, 레오나르 뒤세노, 피에르 주세페 세사, 아칸샤 차우드헤리, 아담 로버츠, 아디티야 바루아, 알렉스 보테프, 알렉스 카스트로-로스, 앰브로스 슬론, 아멜리 엘리우, 안드레아 타케티, 안나 불라노바, 안토니아 패터슨, 베스 차이, 보박 샤리아리, 샤를린 르 랑, 크리스토퍼 A. 쇼케트-추, 클레망 크레피, 다니엘 세르, 다프네 이폴리토, 데이비드 리드, 엘레나 부차츠카야,

에릭 니, 에릭 놀랜드, 겅 얀, 조지 터커, 조지-크리스티안 무라루, 그리고리 로즈데스트벤스키, 헨리크 미할레프스키, 이안 테니, 이반 그리센코, 제이콥 오스틴, 제임스 킬링, 제인 라바노프스키, 장-밥티스트 레스피오, 제프 스탠웨이, 제니 브레넌, 제레미 첸, 요한 페레, 저스틴 치우, 저스틴 마오-존스, 캐서린 리, 캐시 유, 케이티 밀리컨, 라스 로우 쇠순드, 리사 리, 루카스 딕슨, 마셸 리드, 마체이 미쿨라, 마테오 비르트, 마이클 샤먼, 니콜라이 치나예프, 니툼 타인, 올리비에 바셤, 오스카 창, 오스카 발티네즈, 페이지 베일리, 폴 미셸, 페트코 요토프, 라마 차부니, 라모나 코마네스쿠, 리나 자나, 로한 아닐, 로스 맥일로이, 류 루이보, 라이언 멀린스, 새뮤얼 엘스미스, 세바스티앙 보르고, 세르탄 지르긴, 솔토 더글러스, 슈리 판디아, 시아막 샤케리, 소함 데, 테드 클리멘코, 톰 헤니건, 블라드 파인버그, 보이체흐 스토코비에츠, 천유휘, 자파랄리 아메드, 지타오 공, 트리스 워켄틴, 루도비크 페란, 민 장, 클레망 파라베, 오리올 비냐스, 제프 딘, 코라이 카부크추오글루, 데미스 하사비스, 주빈 가라마니, 더글러스 에크, 조엘 바랄, 페르난도 페레이라, 엘리 콜린스, 아르망 쥬랭, 노아 피델, 에반 센터, 알렉 안드레예프, 캐슬린 케닐리. 2024. Gemma: Open models based on gemini research and technology. *Preprint,* arXiv:2403.08295.

세르게이 글로토프. 2023. 문화적 오표현에 맞선 문화간 영화 리터러시 교육: 핀란드 시각예술 교사들의 관점. *미디어 리터러시 교육 저널*, 15(1):31–43.

타오 공, 청치 류, 실롱 장, 유동 왕, 묘 정, 첸 자오, 퀴쿤 리우, 웬웨이 장, 핑 루오, 카이 천. 2023. 멀티모달-gpt: 인간과의 대화를 위한 비전 및 언어 모델. *arXiv 사전 인쇄본 arXiv:2305.04790*.

Harald Hammarström, Robert Forkel, Martin Haspel-math, Sebastian Bank. 2024. Glottolog 5.0. 라이프치히: 막스 플랑크 진화인류학연구소.

타미드 하산, 아비크 바타차르지, Md. 사이풀 이슬람, 카지 무바시르, 위안팡 리, 용빈 강,
M. 소헬 라만, 리파트 샤리야르. 2021. XL-sum: 44개 언어에 대한 대규모 다국어 추상적 요약. 《*계산언어학회 연구 성과: ACL-IJCNLP 2021*》, 4693–4703쪽, 온라인. 계산언어학회.

펑청 허, 지안펑 가오, 웨이주 천. 2022. Debertav3: 기울기 분리 임베딩 공유를 통한 electra 스타일 사전 훈련으로 deberta 개선. *제11회 학습 표현 국제 학회*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Po-Yao Huang, Mandela Patrick, Junjie Hu, Graham Neubig, Florian Metze, Alexander Hauptmann.
2021. 비전-언어 모델의 제로샷 교차언어 전이를 위한 다국어 다중 모달 프리트레이닝. *2021년 북미 컴퓨터 언어학회: 인간 언어 기술 컨퍼런스 논문집*, 2443–2459쪽, 온라인. 컴퓨터 언어학회.

Tin Van Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen. 2022. ViNLI: 개방형 도메인 자연어 추론 연구를 위한 베트남어 코퍼스. *제29회 국제 컴퓨터 언어학 학술대회 논문집*, 3858–3872쪽, 대한민국 경주. 국제 컴퓨터 언어학 위원회.

Muhammad Ichsan. 2023. Merak-7b: 인도네시아어를 위한 대규모 언어 모델. *Hugging Face 저장소*.

조셉 마빈 임페리얼, 제롬 오로스코, 쉴라 메이 마조, 라니 마세다. 2019. 표준 및 양방향 재귀 신경망을 이용한 태풍 관련 트윗 감정 분석. *arXiv 사전 인쇄본 arXiv:1908.01765*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Men-sch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guil-laume Lample, Lucile Saulnier 외. 2023. Mistral 7b. *arXiv 사전 인쇄본 arXiv:2310.06825*.

장성이, 푸시후이, 린난카이, 푸잉원. 2022. 크메르어 사전 훈련 모델 및 평가 데이터. *Tsinghua Science and Technology*, 27(4):709–718.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 자연어 처리 분야에서 언어 다양성과 포용성의 현황과 전망. *제58회 전산언어학회 연차대회 논문집*, 6282–6293쪽, 온라인. 전산언어학회.

사라 삼손 후안, 로랑 베사시에, 벤자민 르쿠투, 모하메드 디압. 2015. 자원이 극히 부족한 언어의 음성인식 개발을 위해 밀접하게 관련된 언어의 자원 활용: 이반어 사례 연구. *INTERSPEECH 학술대회 논문집*, 독일 드레스덴.

Anubha Kabra, Emmy Liu, Simran Khanuja, Al-ham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, Graham Neubig. 2023. 다국어 및 다문화적 비유적 언어 이해. *계산언어학회 연구 성과: ACL 2023,* 8269–8284쪽, 캐나다 토론토. 계산언어학회.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: 인도어용 단일 언어 코퍼스, 평가 벤치마크 및 사전 훈련된 다국어 언어 모델

언어를 위한 *단일 언어* 코퍼스, 평가 벤치마크 및 사전 훈련된 다국어 언어 모델. *계산언어학회 연구 결과: EMNLP 2020*, 4948–4961쪽, 온라인. 계산언어학회.

이촨울 무슬림 카로 카로, 모하마드 파르한 모하마드 푸지, 샤린 카심, 아지줄 아자르 람리. 2022. 기계 학습을 이용한 카로네스어 트윗 감정 분석. *인도네시아 전기 공학 및 정보학 저널(IJEEI)*, 10(1):219–231.

심란 카누자, 세바스찬 루더, 파르타 탈루크다르. 2023. 자연어 처리 기술의 다양성, 형평성 및 포용성 평가: 인도 언어 사례 연구. *컴퓨터 언어학 협회 연구 결과: EACL 2023*, 1763–1777쪽, 크로아티아 두브로브니크. 컴퓨터 언어학 협회.

Fajri Koto, Nurul Aisyah, Haonan Li, Timothy Baldwin. 2023a. 대규모 언어 모델은 인도네시아 초등학교 시험만 통과한다: IndoMMLU에 대한 포괄적 테스트. *2023년 자연어 처리 경험적 방법 컨퍼런스(EMNLP) 논문집*, 싱가포르. Association for Computational Linguistics.

F ajri Koto, Nurul Aisyah, Haonan Li, Timothy Baldwin. 2023b. 대규모 언어 모델은 인도네시아 초등학교 시험만 통과한다: IndoMMLU에 대한 포괄적 테스트. *자연어 처리 경험적 방법론 컨퍼런스 2023 논문집*, 12359–12374쪽, 싱가포르. 계산언어학회.

Fajri Koto, Timothy Baldwin, Jey Han Lau. 2022. 인도네시아어 상식 이야기의 심층적 이해를 위한 클로즈 평가. *상식 표현 및 추론 워크숍(CSRR 2022) 논문집*, 8–16쪽, 아일랜드 더블린. 컴퓨터 언어학회.

Fajri Koto, Ikhwan Koto. 2020. 미낭카바우어의 계산언어학 향상을 위한 연구 : 감정 분석과 기계 번역 연구. *제34회 태평양 아시아 언어 정*보·계산 학회 논문집, 138–148쪽, 베트남 하노이. Association for Computational Linguistics.

파즈리 코토, 라흐마드 마헨드라, 누룰 아이샤, 티모시 볼드윈. 2024. 인도네시아 문화: 11개 인도네시아 주를 아우르는 지리적 영향 하의 문화적 상식 추론 탐구. *사전 인쇄본*, arXiv:2404.01854.

스네하 쿠두군타, 아이작 캐스웰, 비아오 장, 자비에 가르시아, 데릭 신, 아디티야 쿠수파티, 로미 스텔라, 안쿠르 바프나, 오르한 피라트. 2024. Madlad-400: 다국어 및 문서 수준 대규모 검증 데이터셋. *제37회 신경정보처리시스템 국제학술대회(*NIPS '23*) 논문집*, 미국 뉴욕주 레드훅. 커런 어소시에이츠(Curran Associates Inc.)

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, Pratyush Kumar. 2022. IndicNLG 벤치마크: 인도계 언어의 다양한 NLG 작업을 위한 다국어 데이터셋. *2022년 자연어 처리 경험적 방법 컨퍼런스 논문집*, 5363–5394쪽, 아랍에미리트 아부다비. 계산언어학회.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, Victor Sanh. 2024. 비전-언어 모델 구축 시 중요한 요소들. *사전 인쇄본*, arXiv:2405.02246.

Thang Le, Anh Luu. 2023. 베트남 중북부 방언 텍스트 전이를 위한 병렬 코퍼스. *계산언어학회 연구 성과: EMNLP 2023*, 13839–13855쪽, 싱가포르. 계산언어학회.

콜린 리옹, 조슈아 네메체크, 제이콥 만스도퍼, 안나 필리게라, 아브라함 오워둔니, 대니얼 화이트낙. 2022. 블룸 라이브러리: 다양한 다운스트림 작업을 위한 300개 이상의 언어로 된 다중 모달 데이터셋. *2022년 자연어 처리 실증적 방법론 컨퍼런스 논문집*, 8608–8621쪽, 아랍에미리트 아부다비. 계산언어학회.

Wei Qi Leong, Jian Gang Ngui, Yosephine Su-santo, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, William Chandra Tjhi. 2023. Bhasa: 대규모 언어 모델을 위한 동남아시아 언어 및 문화 종합 평가 스위트. *arXiv 사전 인쇄본 arXiv:2309.06085*.

리 하오난, 파즈리 코토, 우 밍하오, 알함 피크리 아지, 티모시 볼드윈. 2023a. Bactrian-x: 저순위 적응을 통한 다국어 재현 가능한 지시 따르기 모델. *arXiv 사전 인쇄본 arXiv:2305.15011*.

Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xu-anjing Huang, and Zhongyu Wei. 2023b. 약한 지도 학습 기반 다국어 시각-언어 사전 훈련을 위한 교차 언어 및 교차 모달 모델링 통합. *제61회 전산언어학회 연례 회의 논문집(제1권: 장문 논문)*, 5939–5958쪽, 캐나다 토론토. 전산언어학회.

Chin-Yew Lin. 2004. ROUGE: 요약 자동 평가를 위한 패키지. 『*텍스트 요약의 분화*』, 74–81쪽, 스페인 바르셀로나. 계산언어학회.

시 빅토리아 린, 토도르 미하일로프, 미켈 아르테체, 왕톈루, 천후이 첸, 다니엘 시미그, 마일 오트, 나만 고얄, 슈루티 보살레, 두징페이, 라마칸트 파수누루, 샘 슐라이퍼, 푸닛 싱 쿠라, 비슈라브 차우다리, 브라이언 오호로, 제프 왕, 루크 제틀-모이어, 조르니차 코자레바, 모나 디아브, 베셀린 스토야노프, 시안 리. 2022. 다국어 생성 언어 모델을 활용한 소량 데이터 학습. *자연어 처리 분야 경험적 방법론 2022년 학술대회 논문집*

9019–9052쪽, 아랍에미리트 아부다비. 계산언어학회.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(1):316.

천 세실리아 류, 파즈리 코토, 티모시 볼드윈, 이리나 구레비치. 2024. 다국어 대규모 언어 모델은 문화적으로 다양한 추론자일까? 다문화 속담과 격언에 대한 연구. *사전 인쇄본*, arXiv:2309.08591.

Haotian Liu, Chunyuan Li, Yuheng Li, Yong Jae Lee. 2023a. 시각적 지시 튜닝을 통한 개선된 기준선.

하오 티안 리우, 춘위안 리, 칭양 우, 용재 리. 2023b. 시각적 지시문 튜닝. *NeurIPS*.

Shayne Longpre, Yi Lu, Joachim Daiber. 2021. MKQA: 다국어 오픈 도메인 질문응답을 위한 언어적으로 다양한 벤치마크. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla 외. 2023. 데이터 출처 이니셔티브: AI 분야의 대규모 데이터셋 라이선싱 및 출처 표시 감사. *arXiv 사전 인쇄본 arXiv:2310.16787*.

일랴 로시칠로프와 프랭크 후터. 2019. 분리된 가중치 감쇠 정규화. *국제 표현 학습 컨퍼런스*.

마누엘 마거, 아르투로 온세바이, 아네트 리오스, 이반 블라디미르 메자 루이즈, 알렉시스 팔머, 그레이엄 노이빅, 카타리나 칸 편집. 2021. *아메리카 원주민 언어 자연어 처리 제1회 워크숍 논문집*. 계산언어학회, 온라인.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, Clara Vania. 2021. IndoNLI: 인도네시아어 자연어 추론 데이터셋. *2021년 자연어 처리 경험적 방법 컨퍼런스 논문집*, 10511–10527쪽, 온라인 및 도미니카 공화국 푼타 카나. 계산언어학회.

니클라스 뮌니호프, 토마스 왕, 린탕 수타위카, 아담 로버츠, 스텔라 비더만, 테벤 르 스카오, M 사이풀 바리, 셩 셴, 정 신용, 하이리 슐쾨프, 샹루 탕, 드라고미르 라데프, 알함 피크리 아지, 칼리드 알무바라크, 사무엘 알바니, 자이드 알야페이, 앨버트 웹슨, 에드워드 래프, 콜린 래펠. 2023. 다중 작업 미세 조정을 통한 교차 언어 일반화. *제61회 전산언어학회 연차 총회 논문집 (제1권: 장문 논문)*, 15991–16111쪽, 캐나다 토론토. 전산언어학회.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf 외. 2022. 다중 작업 미세 조정을 통한 다국어 일반화. *arXiv 사전 인쇄본 arXiv:2211.01786*.

아드 무자드, 파이살 라후토모. 2016. 깊이 우선 집중 크롤링을 통한 인도네시아어 온라인 뉴스 코퍼스. *센트리노프(혁신적 응용 연구 전국 세미나) 논문집*, 2(1):11–20.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 875–880, Brussels, Belgium. Association for Computational Linguistics.

키엣 응우옌, 부 응우옌, 안 응우옌, 응안 응우옌. 2020. 기계 독해력 평가를 위한 베트남어 데이터셋. *제28회 국제 컴퓨터 언어학 학술대회 논문집*, 2595–2605쪽, 스페인 바르셀로나 (온라인). 국제 컴퓨터 언어학 위원회.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. Cul-turaX: 167개 언어의 대규모 언어 모델을 위한 정제되고 방대한 다국어 데이터셋. *2024년 국제 컴퓨터 언어학, 언어 자원 및 평가 공동 학술대회 (LREC-COLING 2024) 논문집*, 4226–4237쪽, 이탈리아 토리노. ELRA 및 ICCL.

Xuan-Phi Nguyen, Wenxuan Zhang, Li Xin, Mahani Aljunied, Weiwen Xu, Hou Pong Chan, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023. Seallms - 동남아시아를 위한 대규모 언어 모델. *사전 인쇄본*, arXiv:arXiv:2312.00738.

오픈AI, 조쉬 아치암, 스티븐 애들러, 산디니 아가르왈, 라마 아마드, 일게 아카야, 플로렌시아 레오니 알레만, 디오고 알메이다, 얀코 알텐슈미트, 샘 알트만, 시얌 알 아나드카트, 레드 아빌라, 이고르 바부쉬킨, 수치르 발라지, 발레리 발콤, 폴 발테스쿠, 하이밍 바오, 모하마드 바바리안, 제프 벨검, 이르완 벨로, 제이크 버딘, 가브리엘 버나뎃-샤피로, 크리스토퍼 버너, 레니 보그도노프, 올렉 보이코, 마들렌 보이드, 안나-루이사 브랙만, 그렉 브록맨, 팀 브룩스, 마일스 브런데이지, 케빈 버튼, 트레버 차이, 로지 캠벨, 앤드류 캔, 브리태니 캐리, 첼시 칼슨, 로리 카마이클, 브룩 찬, 체 창, 포티스 찬치스, 데릭 첸, 슐리 첸, 루비 첸, 제이슨 첸, 마크 첸, 벤 체스, 체스터 조, 케이시 추, 정형원, 데이브 커밍스, 제레미아 커리어, 다이윤싱, 코리 데카로, 토마스 데그리, 노아 도이치, 데미안 드빌, 아르카 다르, 데이비드 도한, 스티브 다울링, 쉴라 더닝, 아드리앙 에코페, 아티 엘레티,

티나 엘룬두, 데이비드 파히, 리암 페더스, 니코 펠릭스, 시몬 포사다 피쉬먼, 저스톤 포르테, 이사벨라 풀포드, 레오 가오, 엘리 조지스, 크리스천 깁슨, 빅 고엘, 타룬 고기네니, 가브리엘 고, 라파 곤티조-로페스, 조너선 고든, 모건 그라프스타인, 스콧 그레이, 라이언 그린, 조슈아 그로스, 시샹 셰인 구, 유페이 구오, 크리스 할라시, 제시 한, 제프 해리스, 위첸 허, 마이크 히튼, 요하네스 하이데케, 크리스 헤세, 앨런 히키, 웨이드 히키, 피터 호에셀레, 브랜든 호튼, 케니 쉬, 셩리 후, 신 후, 요스트 후이징가, 샨타누 제인, 숀 제인, 조앤 장, 안젤라 장, 로저 장, 하오준 진, 데니 진, 시노 조모토, 빌리 존, 희우 준, 토머 카프탄, 루카시 카이저, 알리 카말리, 잉마르 카니체이더, 니티시 시리쉬 케스카르, 타바라크 칸, 로건 킬패트릭, 김종욱, 크리스티나 김, 김용직, 얀 헨드릭 키르히너, 제이미 키로스, 맷 나이트, 다니엘 코코타일로, 루카시 콘드라추크, 앤드류 콘드리치, 아리스 콘스탄티니디스, 카일 코식, 그레첸 크루거, 비샬 쿠오, 마이클 램프, 이카이 란, 테디 리, 얀 라이케, 제이드 렁, 다니엘 레비, 차크 밍 리, 레이첼 림, 몰리 린, 스테파니 린, 마테우스 리트윈, 테레사 로페즈, 라이언 로우, 패트리샤 루, 안나 마칸주, 킴 말파치니, 샘 매닝, 토도르 마르코프, 야니브 마르코프스키, 비앙카 마틴, 케이티 메이어, 앤드류 메인, 밥 맥그루, 스캇 메이어 맥키니, 크리스틴 맥리비, 폴 맥밀란, 제이크 맥닐, 데이비드 메디나, 알록 메타, 제이콥 메닉, 루크 메츠, 안드레이 미셴코, 파멜라 미쉬킨, 비니 모나코, 에반 모리카와, 다니엘 모싱, 통 무, 미라 무라티, 올렉 머크, 데이비드 멜리, 애쉬빈 네어, 레이이치로 나카노, 라지브 나약, 아르빈드 닐라칸탄, 리처드 응오, 현우 노, 롱 우양, 컬런 오키프, 야쿠브 파초키, 알렉스 파이노, 조 팔레르모, 애슐리 판툴리아노, 지암바티스타 파라산돌로, 조엘 패리시, 에미 파르파리타, 알렉스 파소스, 미하일 파블로프, 앤드류 펭, 아담 페렐만, 필리페 데 아빌라 벨부테 페레스, 마이클 페트로프, 엔리케 폰데 드 올리베이라 핀투, 마이클, 포코르니, 미셸 포크라스, 비치르 H. 퐁, 톨리 파월, 알레시아 파워, 보리스 파워, 엘리자베스 프로엘, 라울 푸리, 알렉 래드포드, 잭 레이, 아디티야 라메시, 캐머런 레이먼드, 프랜시스 리얼, 켄드라 림바흐, 칼 로스, 밥 로트스테드, 앙리 루세, 닉 라이더, 마리오 살타렐리, 테드 샌더스, 시바니 산투르카르, 기리쉬 사스트리, 헤더 슈미트, 데이비드 슈너, 존 슐만, 다니엘 셀삼, 카일라 셰퍼드, 토키 셰르바코프, 제시카 시에, 사라 쇼커, 프라나브 샤얌, 시몬 시도르, 에릭 시글러, 매디 시멘스, 조던 시트킨, 카타리나 슬라마, 이안 솔, 벤자민 소콜로프스키, 양 송, 나탈리 스타우다허, 펠리페 페트로스키 수크, 나탈리 서머스, 일리아 수츠케버, 지에 탕, 니콜라스 테작, 매들린 B. 톰슨, 필 틸레, 아민 투툰치안, 엘리자베스 쳉, 프레스턴 터글, 닉 터리, 제리 트워렉, 후안 펠리페 세론 우리베, 안드레아 발론, 아룬 비자이베르기야, 첼시 보스, 캐럴 웨인라이트, 저스틴 제이 왕, 앨빈 왕, 벤 왕, 조너선 워드, 제이슨 웨이, CJ 바인만, 아킬라 웰린다, 피터 웰린더, 지아이 웡, 릴리안 웡, 맷 위트호프, 데이브 윌너, 클레멘스 윈터, 새뮤얼 울리치, 한나 웡, 로렌 워크맨, 셔윈 우, 제프 우, 마이클 우, 카이 샤오, 타오 쉬, 사라 유, 케빈 유, 치엠-

잉 위안, 보이체흐 자렘바, 로완 젤러스, 종 장, 마빈 장, 성자 자오, 천하오 정, 준탕 장, 윌리엄 주크, 바렛 조프. 2024. GPT-4 기술 보고서. *사전 인쇄본*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray 외. 2022. 인간 피드백을 통한 지시사항 준수 언어 모델 훈련. *신경정보처리시스템 발전*, 35:27730–27744.

체스터 팔렌-미셸, 콘스탄틴 리그노스. 2023. LR-sum: 자원이 부족한 언어를 위한 요약. *컴퓨터 언어학 협회 연구 결과: ACL 2023*, 6829–6844쪽, 캐나다 토론토. 컴퓨터 언어학 협회.

Wannaphong Phatthiyaphaibun, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkon-chotiwat, Thanathip Suntorntip, Can Udom-charoenchaikit. 2023. PyThaiNLP: Python 기반 태국어 자연어 처리. *제3회 자연어 처리 오픈소스 소프트웨어 워크숍(NLP-OSS 2023) 논문집*, 25–36쪽, 싱가포르. Association for Computational Linguistics.

완나퐁 파티야파이분, 수라폰 노네순, 파톰폰 파융캄디, 피랏 림콘초티왓, 칸우돔차로엔차이킷, 짠카팟 사왓폴, 촘파콘 차크상차이촙, 에카폴 추앙수완니치, 사라나 누타농. 2024. 왕찬리온과 왕찬엑스 mrc 평가. *사전 인쇄본*, arXiv:2403.16127.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulic´, Anna Korhonen. 2020. XCOPA: 인과적 상식 추론을 위한 다국어 데이터 세트. *2020년 자연어 처리의 경험적 방법 컨퍼런스(EMNLP) 논문집*, 2362–2376쪽, 온라인. 컴퓨터 언어학 협회.

마야 포포비치. 2015. chrF: 자동 기계 번역 평가를 위한 문자 n-그램 F-점수. *제10회 통계적 기계 번역 워크숍 논문집*, 392–395쪽, 포르투갈 리스본. 계산언어학회.

Ma j a Popovic´. 2017. chrF++: 문자 n-그램을 지원하는 단어. *제2회 기계 번역 컨퍼런스 논문집*, 612–618쪽, 덴마크 코펜하겐. 계산언어학회.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, Michael Auli. 2024. 1,000개 이상의 언어로 확장되는 음성 기술. *기계 학습 연구 저널*, 25(97):1–52.

The Joshua Project. 2024. The joshua project.

Ayu Purwarianti 및 Ida Ayu Putu Ari Crisdayanti. 2019. 단락 벡터를 사용한 인도네시아어 감정 분석을 위한 bi-lstm 성능 개선. *2019 국제 고급 정보학 컨퍼런스: 개념, 이론 및 응용 (ICAICTA)*, 1–5쪽. IEEE.

A yu Purwarianti, Masatoshi Tsuchiya, and Seiichi Nak-agawa. 2007. 인도네시아어 질문응답 시스템을 위한 기계 학습 접근법. *인공 지능 및 응용*, 573–578쪽.

I Made Suwija Putra, Daniel Siahaan, Ahmad Saikhu. 2024. SNLI Indo: 스탠퍼드 자연어 추론 데이터셋에서 파생된 인도네시아어 텍스트 함축 인식 데이터셋. *Data in Brief*, 52:109998.

알렉 래드포드, 김종욱, 크리스 할라시, 아디티야 라메시, 가브리엘 고, 산디니 아가르왈, 기리쉬 사스트리, 아만다 애스켈, 파멜라 미쉬킨, 잭 클라크, 그레첸 크루거, 일리아 수츠케버. 2021. 자연어 감독을 통한 전이 가능한 시각 모델 학습. *제38회 국제 기계 학습 컨퍼런스 논문집, 기계 학습 연구 논문집* 제139권, 8748–8763쪽. PMLR.

알렉 래드포드, 김종욱, 타오 쉬, 그렉 브록맨, 크리스틴 맥리비, 일리아 수츠케버. 2023. 대규모 약한 지도 학습을 통한 견고한 음성 인식. *제40회 국제 기계 학습 학회 논문집, 기계 학습 연구 논문집* 제202권, 28492–28518쪽. PMLR.

리코산, 카렌 에타니아 사푸트라. 2023. 인도네시아 모바일 애플리케이션 리뷰 기반 다중 레이블 다중 클래스 감정 및 정서 데이터셋. *데이터 인 브리프*, 50:109576.

Parker Riley, Isaac Caswell, Markus Freitag, David Grangier. 2020. "다국어" NMT에서 번역어로서의 번역어. *제58회 전산언어학회 연차대회 논문집*, 7737–7746쪽, 온라인. 전산언어학회.

무함마드 라지프 리즈쿨라, 아유 푸르와리안티, 알함 피크리 아지. 2023. 카시나: 시라 나바위야를 활용한 종교 분야 질문 답변 시스템. *2023 제10회 고급 정보학 국제 컨퍼런스: 개념, 이론 및 응용 (ICAICTA)*, 1–6쪽. IEEE.

앤드류 루디첸코, 사미르 쿠라나, 사무엘 토마스, 로제리오 페리스, 레오니드 칼린스키, 힐데 퀴네, 데이비드 하르바스, 브라이언 킹스버리, 제임스 글래스. 2023. 다국어 자기지도 및 약지도 음성 사전 훈련의 미관측 언어 적응 비교. *INTERSPEECH 2023 논문집*, 2268–2272쪽.

Sebastian Ruder, Jonathan H Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijh-wani, Parker Riley, Jean-Michel Sarr, Xinyi Wang 외. 2023. Xtreme-up: 사용자 중심의 저조한 데이터 벤치마크를 위한 저조한 언어. *Find-ings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1856–1884.

빅터 산, 앨버트 웹슨, 콜린 라펠, 스티븐 H. 바흐, 린탕 수타위카, 자이드 알야페이, 앙투안 샤핀, 아르노 스티글러, 테벤 르 스코, 아룬 라자, 마난 데이, 엠 사이풀 바리, 캉웬 쉬, 우르미쉬 타커, 샤냐 샤르마 샤르마, 엘리자 슈체클라, 김태운, 군잔 차블라니, 니할 나약, 데바조티 다타, 조너선 창, 마이크 티안-지안 장, 한 왕, 마테오 마니카, 셩 센, 정신용, 하르시트 판데이, 레이첼 바우든, 토마스 왕, 트리샬라 니라즈, 조스 로젠, 아비슈트 샤르마, 안드레아 산틸리, 티보 페브리, 제이슨 앨런 프라이스, 라이언 티한, 스텔라 비더만, 레오 가오, 탈리 버스, 토마스 울프, 알렉산더 M. 러시. 2021. 다중 작업 프롬프트 훈련을 통한 제로샷 작업 일반화. *사전 인쇄본*, arXiv:2110.08207.

A uliya Sani, Sakriani Sakti, Graham Neubig, Tomoki Toda, Adi Mulyanto, and Satoshi Nakamura. 2012. 언어 보존을 향하여: 인도네시아 민족 언어 데이터의 예비 수집 및 모음 분석. *2012 국제 음성 데이터베이스 및 평가 컨퍼런스*, 118–122쪽.

크리스토프 슈만, 로맹 보몽, 리처드 벤쿠, 케이드 고든, 로스 와이트먼, 메디 체르티, 테오 쿰브스, 아루쉬 카타, 클레이튼 멀리스, 미첼 워츠만 외. 2022. Laion-5b: 차세대 이미지-텍스트 모델 훈련을 위한 공개 대규모 데이터셋. *신경정보처리시스템 발전*, 35:25278–25294.

켄 나빌라 세티아와 라흐마드 마헨드라. 2018. 인도네시아어 위키피디아 데이터에 대한 반감독 학습 텍스트 함축. *국제 컴퓨터 언어학 및 지능형 텍스트 처리 학회*, 416–427쪽. Springer.

AI Singapore. 2023. Sea-lion (southeast asian languages in one network): 동남아시아를 위한 대규모 언어 모델 패밀리. https://github.com/aisingapore/ sealion.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony 외. 2024. Aya 데이터셋: 다국어 지시어 튜닝을 위한 오픈 액세스 컬렉션. *arXiv 사전 인쇄본 arXiv:2402.06619*.

앤더스 소가드. 2022. 영어 *자연어*처리(NLP)를 1년간 금지해야 할까? 《*2022년 자연어 처리 경험적 방법론 학회 논문집*》, 5254–5260쪽, 아부다비, 아랍에미리트. 계산언어학회.

Rhio Sutoyo, Said Achmad, Andry Chowanda, Es-ther Widhi Andangsari, Sani M. Isa. 2022.

Prdect-id: 감정 분류 작업을 위한 인도네시아 제품 리뷰 데이터셋. *Data in Brief*, 44:108554.

Oscar Täckström, Ryan McDonald, Jakob Uszkor-eit. 2012. 언어 구조의 직접적 전이를 위한 교차 언어 단어 클러스터. *2012년 북미 컴퓨터 언어학 협회 컨퍼런스: 인간 언어 기술* 회의록, 477–487쪽, 캐나다 몬트리올. 컴퓨터 언어학 협회.

제락 탈라트, 오렐리 네베올, 스텔라 비더만, 미루나 클린치우, 마난 데이, 셰인 롱프레, 사샤 루크시오니, 마라이임 마수드, 마거릿 미첼, 드라고미르 라데프, 샤냐 샤르마, 아르준 수브라모니안, 태재성, 삼손 탄, 디팍 투누군틀라, 오스카 반 데르 발. 2022. 뿌린 대로 거두리라: 다국어 환경에서의 편향 평가 과제. *『BigScience 에피소드 #5 – 대규모 언어 모델 구축의 과제 및 전망 워크숍』 논문집*, 26–41쪽, 가상+더블린. 계산언어학회.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: 대규모 다국어 다중 모달 평가 데이터셋. *2022년 자연어 처리 경험적 방법 컨퍼런스 논문집*, 715–729쪽, 아랍에미리트 아부다비. 컴퓨터 언어학 협회.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar 외. 2023. Llama: 개방적이고 효율적인 기초 언어 모델. *arXiv 사전 인쇄본 arXiv:2302.13971*.

Khanh Quoc Tran, Phap Ngoc Trinh, Khoa Nguyen-Anh Tran, An Tran-Hoai Le, Luan Van Ha, Kiet Van Nguyen. 2021. 베트남어 대규모 고품질 오픈 도메인 데이터셋을 활용한 온라인 뉴스 분류 실증 연구. *New Trends in Intelligent Software Methodologies, Tools and Techniques*, pp. 367–379. IOS Press.

아흐메트 우스툰, 비라트 아리아부미, 정신용, 고웨이인, 다니엘 드수자, 그베밀레케 오닐루데, 닐 반다리, 시발리카 싱, 우이리 우이, 암르 카이드 외. 2024. Aya 모델: 지시어 미세조정이 적용된 오픈 액세스 다국어 언어 모델. *arXiv 사전 인쇄본 arXiv:2402.07827*.

키엣 반 응우옌, 틴 반 휩, 덕-부 응우옌, 안 지아-투안 응우옌, 응안 루-투이 응우옌. 2022. 건강 뉴스 기사 기계 독해 이해를 위한 새로운 베트남어 코퍼스. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(5).

R amakrishna Vedantam, C Lawrence Zitnick, Devi Parikh. 2015. Cider: 합의 기반 이미지 설명 평가. *IEEE 컴퓨터 비전 및 패턴 인식 학회 논문집*, 4566–4575쪽.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, Nancy F Chen. 2023. 다국어 기초 모델을 위한 Seaeval: 교차 언어 정렬에서 문화적 추론까지. *arXiv 사전 인쇄본 arXiv:2309.04766*.

왕빈, 류정위안, 황신, 교팡카이, 딩양, 아이티 아우, 낸시 F. 첸. 2024. 다국어 기초 모델을 위한 Seae-val: 교차 언어 정렬에서 문화적 추론까지. *NAACL*.

제이슨 웨이, 마르텐 보스마, 빈센트 Y 자오, 켈빈 구, 애덤스 웨이 유, 브라이언 레스터, 난 두, 앤드류 M 다이, 콴 V 레. 2021. 미세 조정된 언어 모델은 제로샷 학습자이다. *arXiv 사전 인쇄본 arXiv:2109.01652*.

하리오 아크바리안토 위보워, 에를란드 힐만 푸아디, 마데 닌디야타마 니티아샤, 라디티오 에코 프라소조, 알함 피크리 아지. 2023. Copal-id: 지역 문화와 미묘한 차이를 고려한 인도네시아어 추론. *arXiv 사전 인쇄본 arXiv:2311.01012*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, Ayu Purwarianti. 2020. I n d oNLU: 인도네시아어 자연어 이해 평가를 위한 벤치마크 및 자원. *제1회 아시아태평양계산언어학회(ACL) 및 제10회 국제자연어처리학술대회(IJNLP) 논문집*, 843–857쪽, 중국 쑤저우. 계산언어학회(ACL).

겐타 인드라 위나타, 알함 피크리 아지, 사무엘 카야위-자야, 라흐마드 마헨드라, 파즈리 코토, 아데 로마드-호니, 케말 쿠르니아완, 데이비드 뫼르자디, 라디-티오 에코 프라소조, 파스칼 펑, 티모시 볼드윈, 제이 한 라우, 리코 센리히, 세바스찬 루더. 2023. Nu s a X: 10개 인도네시아 지역 언어를 위한 다국어 병렬 감정 데이터셋. *제17회 유럽계 컴퓨터언어학회(ECAL) 학술대회 논문집*, 815–834쪽, 크로아티아 두브로브니크. 컴퓨터언어학회(ACL).

Genta Indra Winata, Ruochen Zhang, David Ife-oluwa Adelani. 2024. Miners: 의미 검색기로서의 다국어 언어 모델. *arXiv 사전 인쇄본 arXiv:2406.07424*.

BigScience 워크숍, 테벤 르 스카오, 안젤라 판, 크리스토퍼 아키키, 엘리 파블릭, 수자나 일리치, 다니엘 헤슬로, 로만 카스타뉴, 알렉산드라 사샤 루시오니, 프랑수아 이본 외. 2022. Bloom: 176b 매개변수 오픈 액세스 다국어 언어 모델. *arXiv 사전 인쇄본 arXiv:2211.05100*.

린팅 쉬, 노아 콘스탄트, 아담 로버츠, 미히르 칼레, 라미 알-르푸, 아디티야 시단트, 아디티야 바루아, 콜린 라펠. 2021. mT5: 대규모 다국어 사전 훈련된 텍스트-투-텍스트 트랜스포머. *2021년 북미 컴퓨터 언어학회 학술대회 논문집:*

인간 언어 기술, 483–498쪽, 온라인. 계산언어학회.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Rowena Garcia, Thamar Solorio, Alham Aji. 2023. 다국어 대규모 언어 모델을 활용한 코드 혼합 텍스트 생성: 동남아시아 언어 사례. *제6회 언어학적 코드 전환에 대한 계산적 접근 워크숍 논문집*, 43–63쪽, 싱가포르. 계산언어학회.

장루오천, 사무엘 카야위자야, 얀 크리스티안 블레이즈 크루즈, 겐타 위나타, 알함 아지. 2023a. 다국어 대규모 언어 모델은 (아직) 코드 스위칭을 하지 않는다. *2023년 자연어 처리 경험적 방법론 컨퍼런스 논문집*, 12567–12582쪽, 싱가포르. 컴퓨터 언어학 협회.

장원쑤안, 마하니 알주니드, 창 가오, 유켄 치아, 빙리동. 2023b. M3exam: 대규모 언어 모델 평가를 위한 다국어, 다중 모달, 다단계 벤치마크. *신경정보처리시스템 발전*, 제36권, 5484–5505쪽. 커런 어소시에이츠.

장원쑤안, 마하니 알주니드, 가오창, 치아위켄, 빙리동. 2024. M3exam: 대규모 언어 모델 평가를 위한 다국어, 다중 모달, 다단계 벤치마크. *신경정보처리시스템 발전*, 36권.

주더야오, 천준, 션샤오치안, 리샹, 엘호세이니 모하메드. 2024. Minigpt-4: 고급 대규모 언어 모델을 통한 시각-언어 이해력 향상. *ICLR*.

SEACrowd의 주요 결론

주요 결과는 다음과 같습니다:

모델 성능.

- **LLM:** AYA-101 및 mT0와 같은 동남아시아(SEA) 전용 모델은 제로샷 작업에서 강력한 성능을 보이며, 해당 지역에서 영어 또는 국가별 모델을 능가합니다. 그러나 추상적 QA 및 요약과 같은 작업에서는 기존 모델이 SEA 언어를 효과적으로 처리하는 데 한계가 드러납니다.
- **음성:** Whisper v3 같은 기성 모델은 주요 동남아시아 언어에서 경쟁력 있는 ASR 성능을 보이지만 토착 언어에서는 어려움을 겪습니다. 반면 Seamless M4T v2는 동남아시아 언어 전반에 걸쳐 더 균형 잡힌 결과를 제공합니다.
- **VLMs:** 현재의 VLM은 동남아시아 언어에서 고품질 이미지 캡션을 생성하지 못하여, 보다 효과적인 다국어 사전 훈련의 필요성을 강조합니다.

LLM 생성 품질. LLM의 동남아시아 언어 출력은 종종 번역체로 인해 문제가 발생하며, SEA-LION v1과 같은 모델은 57.71%의 경우에만 자연스러운 문장을 생성합니다. 타갈로그어, 버마어, 말레이어와 같은 언어는 부자연스러운 생성에 시달립니다.

자원 격차. SEACrowd는 동남아시아 언어의 74.9%를 커버하지만, 대부분의 언어에 포괄적인 데이터셋이 부족한 롱테일 분포를 보여줍니다. 동남아시아 언어는 또한 문화적 왜곡 문제에 직면해 있으며, 데이터셋의 70%가 문화적 관련성이 있는 원본 자료가 아닌 번역본입니다.

개발 우선순위 설정. 자연스러움 측면에서 상당한 격차가 있는 동남아시아 국가 언어(예: 말레이어, 버마어, 필리핀어)와 자바어, 세부아노어와 같은 자원이 부족한 지역 언어에 초점을 맞춰야 합니다.

협력. 정부, 산업계, 지역 사회는 자원 형평성 문제를 해결하고 동남아시아 AI 개발을 개선하기 위해 연구 개발, 데이터 수집 및 개방형 협력에 투자해야 합니다.

B 관련 연구

동남아시아 데이터 자원 동남아시아 언어에 대한 대규모 언어 모델 연구는 이용 가능한 데이터셋과 벤치마크의 부족으로 제한된다. 현재까지 동남아시아 자연어 처리 작업에 대한 자원은 상대적으로 자원이 풍부한 동남아시아 원주민 언어에 집중되어 있다.

벤치마크	언어 수	# 동남아시아 원주민 언어	# 데이터셋	# 작업
SEACrowd (**우리의**)†	39	38	254	13 (11 텍스트, 1 음성, 1 비전)
누사크라우드† (Cahyawijaya 외, 2023a)	19	19	137	12 (11개 텍스트, 1개 연설)
BUFFET (Asai 외, 2023)	54	해당 없음	15	8 (8개 텍스트)
XTREME-UP (Ruder 외, 2023)	88	11	269	9 (텍스트 7개, 음성 1개, 비전 1개)

표 3: 벤치마크 비교. †SEACrowd 및 NusaCrowd의 숫자는 평가에 포함된 데이터셋 수를 나타냅니다.

인도네시아어(Mahendra 외, 2021; Wilie 외, 2020; Cahyawijaya 외, 2021, 2023a) 및 베트남어(Nguyen 외, 2020; Huynh 외, 2022; Le and Luu, 2023; Van Nguyen et al., 2022). NusaCrowd(Cahyawijaya et al., 2023a)는 텍스트와 음성을 포함한 인도네시아어 최초의 다중 모달 벤치마크를 소개한다. Ruder et al.(2023)은 동남아시아(SEA)의 11개 토착 언어를 아우르는 다중 모달 벤치마크를 소개하며, 총 88개에 달하는 다양한 언어를 포괄한다.

또한 아사이 외(2023)는 다양한 다국어 데이터셋에서 추출한 15개의 서로 다른 작업과 54개 언어로 구성된 다국어 소량 데이터 전이 학습을 위한 대규모 언어 모델 벤치마크를 제시한다. 더불어 두 외(2024)는 동남아시아 언어에 대한 공개 사전 훈련 데이터가 텍스트 중복 및 유니코드 이스케이프 과다 발생과 같은 품질 문제점을 안고 있음을 발견했다. 반면 동남아시아 언어 전용으로 사전 훈련된 LLM은 언어 지원 범위가 제한적입니다. 예를 들어 Cendol(Cahyawijaya et al., 2024b), Sailor(Dou et al., 2024), SEA-LION(Singapore, 2023), SeaLLMs(Nguyen et al., 2023)는 영어와 중국어를 포함해 최대 11개 동남아시아 언어만 다루고 있습니다.

NLP 분야의 오픈소스 커뮤니티 이니셔티브 오픈소스 및 오픈사이언스 커뮤니티는 원어민을 참여시켜 대규모 다국어 NLP 자원을 큐레이션하는 데 핵심적인 역할을 합니다. 과거에는 글로벌 규모(Workshop et al., 2022; Singh et al., 2024; Üstün et al., 2024) 또는 지역 수준(예: 아프리카 언어용 Masakhane (Ade-lani et al., 2021, 2022b,a, 2023), 인도 언어용 AI4Bharat (Kakwani et al., 2020; Kumar 외, 2022; Dabre 외, 2022 등), 라틴 아메리카 언어용 Americas-NLP(Mager 외, 2021; Ebrahimi 외, 2022) 등이 있다.

동남아시아 지역에서는 인도네시아어 자연어 처리 연구를 위한 커뮤니티 기반 이니셔티브(예: IndoNLP, PyThaiNLP, RojakNLP)가 진행되어 왔다(Aji et al., 2022; Wilie et al., 2020; Cahyawijaya et al., 2021, 2023a), 태국어(Phatthiyaphai-bun et al., 2023), 그리고 동남아시아(SEA)에서의 코드 스위칭 현상(Aji et al., 2023; Yong et al., 2023; Winata et al., 2024)을 각각 다루었다.

제출	점수	최대 점수
공개 데이터 시트	2+보너스	6
데이터 로더	3	어려울 경우 6
비공개 데이터 시트	1	-
비공개 데이터 접근	4+보너스	고품질 시 10
데이터시트 검토	1	1
데이터 로더 검토	2	어려울 경우 4
비공개 데이터시트 검토	0.5	-
비공개 데이터 문의	1	성공 시 5점

표 4: 데이터시트, 데이터 로더 및 개인 데이터 관련 기여에 대한 획득 포인트 수.

C SEACrowd 기여

C.1 공개 기여

SEACrowd의 공개 기여를 위한 네 가지 과제를 제시합니다.[12] 이 과제들과 SEACrowd의 작업 흐름은 인도네시아어 NLP를 위한 데이터 자원 통합 협력 프로젝트인 NusaCrowd(Cahyawijaya et al., 2023a, 2022)의 영향을 크게 받고 이를 확장한 것입니다.

- **기존 공개 데이터셋에 대한 메타데이터 제출.** 기여자는 이 양식을 통해 기존 데이터셋에 대한 상세한 데이터 시트를 제출할 수 있습니다.[13] 기여자는 데이터 라이선스, 크기, 언어 및 방언, 주석 방법 등과 같은 중요한 정보를 제공해야 합니다. 승인된 데이터 시트와 검토 중인 데이터 시트는 모니터 스프레드시트와 SEACrowd 카탈로그(그림 7)에 표시되고 색인됩니다.
- **데이터 로더** 구축. 이전 작업에서 승인된 데이터시트를 바탕으로, 기여자들은 Hugging Face 데이터셋 로더를 구축하여 모든 데이터셋이

[12] 랜딩 페이지: https://github.com/SEACrowd.

[13] 공개 데이터시트 양식: https://form.jotform.com/team/232952680898069/seacrowd-sea-datasets.

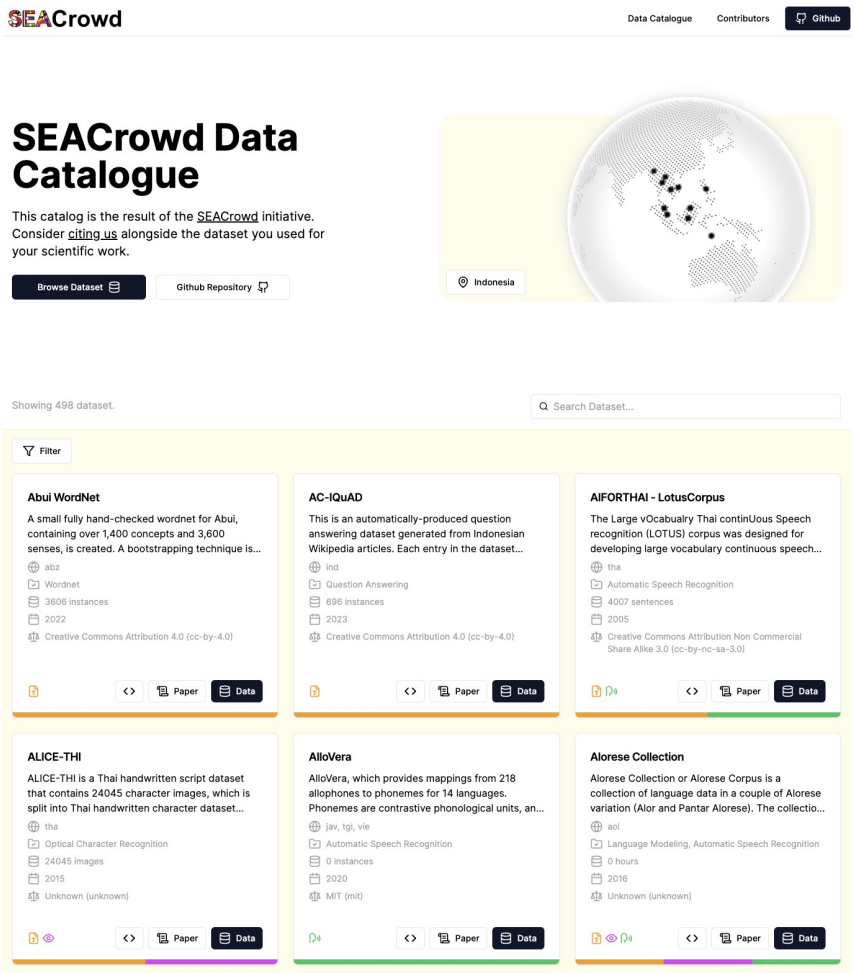


그림 7: SEACrowd 카탈로그의 모습.

SEACrowd의 모든 데이터 세트가 형식 및 사용 측면에서 표준화되도록 보장합니다. 기여자는 SEACrowd 데이터 허브에서 제공되는 데이터 로더 가이드 및 예시를 따를 수 있습니다[14]. 데이터 로더 관리자와 검토자는 또한 2주 동안 활동이 없는 자체 할당된 데이터 로더 문제를 모니터링하고, 차단 장애가 있는 경우 기여자에게 알립니다.

- **동남아시아 언어, 문화 및/또는 지역을 위한 비공개 AI 데이터셋 식별.** 안타깝게도 동남아시아 언어를 다루는 다수의 기존 연구는 여전히 공개되지 않은 상태입니다. 이는 자금 지원과 관련된 비공개 계약, 사적·개인 식별 정보 포함, 영리 기업이 사용하는 것과 같은 명시적 비공개 데이터 활용 등 여러 가지 이유(이에 국한되지 않음)로 인해 발생할 수 있습니다.

  본 작업에서 기여자는 비공개 데이터를 포함하는 연구를 검색하고 해당 정보를 입력할 수 있습니다. 해당 기록 양식을 작성할 수 있습니다.[15] SEACrowd 팀은 이후 원본 데이터 소유자에게 연락하여 자원의 오픈소스화를 협의합니다.

- **SEA의 비공개 AI 데이터셋 공개.** 기여자가 비공개 데이터로 작업한 이력이 있거나(또는 SEACrowd 팀으로부터 비공개 소스 데이터 관련 연락을 받은 경우), 공개 데이터시트 양식을 통해 자원을 공개하고 컬렉션에 등록할 수 있습니다. 해당 리소스는 여전히 원본 기여자의 소유이며 기여자의 기존 작업과 연결됩니다. SEACrowd는 단순히 이를 목록화하고 오픈소스 라이선스로 전환된 상태를 기록할 뿐입니다.

[14] 데이터 로더 가이드: https://github.com/SEACrowd/ seacrowd-datahub/blob/master/DATALOADER.md.

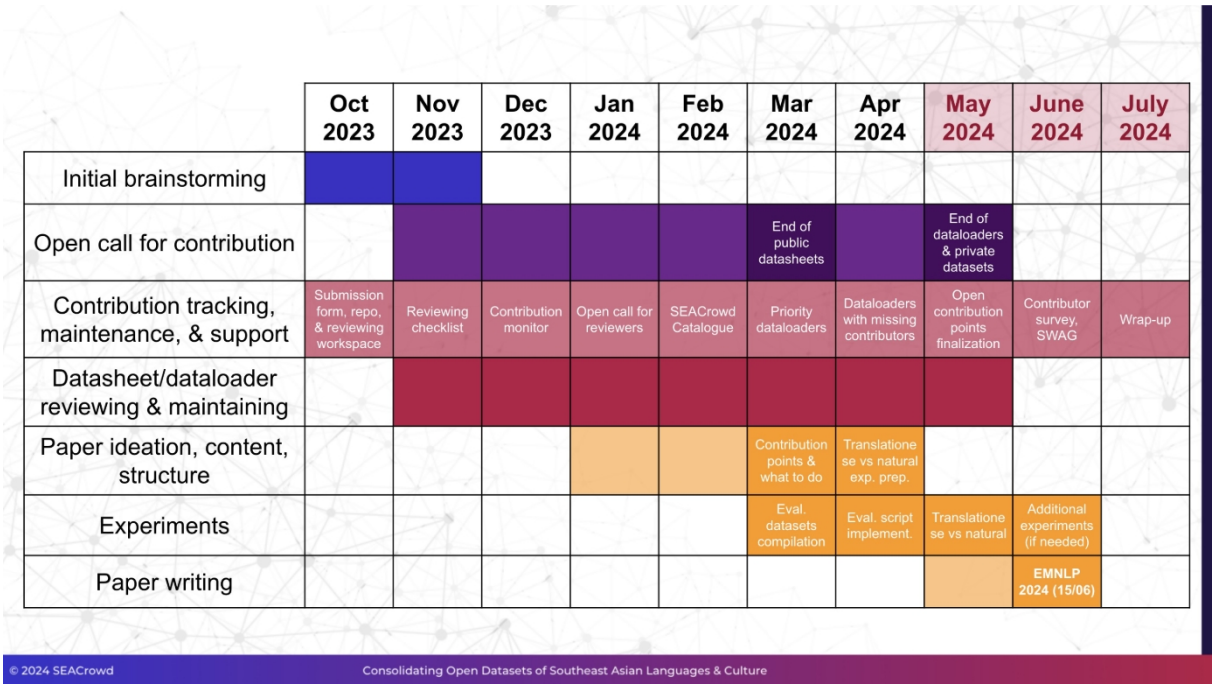


그림 8: SEACrowd 전체 운영 기간의 타임라인.

C.2 기여도 측정

공동 저자로 인정받으려면 20점의 기여도가 필요합니다.[16] 기여자들이 획득한 점수를 모니터링하기 위해 기여도 추적 시스템이 제공되며 정기적으로 업데이트됩니다. 이 점수 제도의 목적은 협력을 방해하기 위한 것이 아니라 희귀하고 고품질의 데이터셋 등록을 보상하기 위함입니다. 표 4는 기여도 점수를 설명합니다.[17] 데이터셋 모달리티가 음성 또는 비전인 경우 1점의 보너스가 부여됩니다. 또한 Joshi 등(2020)([18])이 정의한 가용 자원 측면의 언어 희귀도에 기반한 보너스를 제공합니다.레벨 1 및 2언어는 1점, 레벨 0 또는 목록에 없는 언어는 2점이 부여됩니다. 표 4에 언급되지 않은 기타 기여(예: 유지보수, 설계, 실험, 논문 작성 등)의 경우, 관련 작업의 규모와 복잡성에 따라 기여 점수를 조정합니다.

[15] 논문 with 비공개 데이터셋 형식: https://form.jotform.com/team/232952680898069/seacrowd- paper-with-private-dataset.

[16] 마감일(부록 D.1 참조) 이후 제출된 자료는 여전히 기록되지만, 기여 점수는 더 이상 부여되지 않습니다.

[17] 기여도 점수 가이드라인: https://github.com/ SEACrowd/seacrowd-datahub/blob/master/POINTS.md.

[18] https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt

D SEACrowd의 진행 상황

D.1 타임라인

SEACrowd는 2023년 11월 1일 기여자 공개 모집을 시작했습니다. 데이터시트 제출은 2024년 3월 31일까지, 데이터 로더 및 비공개 데이터셋 하위 제출은 2024년 5월 15일까지 진행되었습니다. SEACrowd 기여자들은 기여 과정에서 직면한 과제, 진행을 위한 다음 단계, 논문 실험 및 연구 아이디어에 대해 격주로 논의합니다. 상세 일정은 그림 8에서 확인할 수 있습니다.

D.2 기여 진행 상황

그림 9는 SEACrowd에서 공개 데이터시트, 데이터 로더 풀 리퀘스트, 비공개 데이터셋을 포함한 논문 제출 건수를 보여줍니다.

E SEACrowd 제출물 검토

데이터 허브에서 완전한 검토 가이드라인을 제공합니다.[19]

E.1 데이터시트 검토

데이터시트 검토 표준 운영 절차(SOP)는 SEACrowd에 제출된 데이터 세트의 무결성과 완전성을 보장합니다. 이 절차는 데이터 세트 가용성 확인 절차,

[19] 검토자 SOP: https://github.com/SEACrowd/ seacrowd-datahub/blob/master/REVIEWING.md

중복 방지, 정확성 확보 및 SEA 지역 관련성 확인을 위한 절차를 명시합니다. 본 SOP에는 데이터셋 중복 및 오류 정보와 같은 일반적인 문제에 대한 FAQ와 데이터 가용성, 데이터셋 분할, 라이선싱 등을 다루는 승인 체크리스트가 포함됩니다. 검토자는 오류 수정 및 다중 기여자에 대한 점수 배정 결정 등 다양한 시나리오 처리 방법을 안내받습니다. 예를 들어 제출된 데이터 시트에 오류나 누락된 정보가 있을 경우, 검토자는 기여자에게 수정 요청(일부 안내 제공)하거나 직접 수정할 수 있습니다. 검토 완료 시 검토자는 상태를 업데이트하고, 메모와 점수를 추가한 후 승인된 데이터 시트에 대한 GitHub 이슈 생성을 기다립니다.

E.2 데이터로더 검토

데이터로더 검토 표준 운영 절차(SOP)는 SEACrowd 내 데이터로더 검토 프로세스를 규정하여 데이터 구조, SEACrowd 스키마 및 구성 표준 준수를 보장합니다. 메타데이터 정확성, 하위 집합 구현, 테스트 스크립트 통과 여부, 코딩 규칙 준수 여부에 대한 검사를 명시합니다. 또한 데이터셋 유형에 따른 데이터 로더 구성 규칙을 제시하고 다국어 데이터셋에 대한 가이드라인을 제공합니다. 본 SOP는 검토자 협업의 중요성을 강조하며, 제출된 각 풀 리퀘스트당 두 명의 검토자가 필요함을 명시합니다. 또한 가용성과 신속성에 기반한 할당 또는 자가 할당 방식의 승인 및 검토자 배정 절차를 규정합니다.

F SEACrowd의 스키마

스키마는 데이터 로더가 반환하는 데이터 세트의 속성을 정의하고 형식을 지정합니다. 각 데이터 로더에 대해 두 가지 유형의 스키마를 구현합니다: 소스 스키마와 seacrowd 스키마입니다. 소스 스키마는 데이터 세트를 원래 구조와 유사한 형식으로 제시하는 반면, seacrowd 스키마는 유사한 작업 전반에 걸쳐 데이터 구조를 표준화합니다. 다음 하위 섹션에서는 NLP(F.1), 음성(F.2), VL(F.3) 분야의 시크라우드 스키마를 정의합니다.

F.1 NLP

- **라벨링되지 않은 텍스트(SSP)**. 이 스키마는 자기 지도 사전 훈련에서의 언어 모델링에 사용될 수 있습니다. (id, text)로 구성되며, id는 데이터셋의 고유 행 식별자를, text는 입력 텍스트를 나타냅니다.
- **단일 레이블 텍스트 분류(TEXT)**. 이 스키마는 감성 분석에 사용될 수 있습니다.

하위 집합 ID	언어	지역	# 샘플
감정 분석 → *_seacro		wd_text	
lazada_review_filipino	FIL	필리핀	1001
gklmip_sentiment	MYA	미얀마	716
인도르_감정	IND	인도네시아	1011
id_sentiment_analysis	IND	인도네시아	10806
karonese_sentiment	BTX	인도네시아	1000
wisesight_thai_sentiment	THA	태국	2671
wongnai_reviews	THA	태국	6203
typhoon_yolanda_tweets	필리핀	필리핀	153
smsa	IND	인도네시아	500
prdect_id_sentiment	IND	인도네시아	5400
id_sent_emo_mobile_apps_sentiment	IND	인도네시아	21696
shopee_reviews_tagalog	FIL	필리핀	2250
nusatranslation_senti_abs	ABS	인도네시아	500
nusatranslation_senti_btk	BTX	인도네시아	1200
nusatranslation_senti_bew	BEW	인도네시아	1200
nusatranslation_senti_bhp	BHP	인도네시아	500
nusatranslation_senti_jav	JAV	인도네시아	1200
nusatranslation_senti_mad	MAD	인도네시아	1200
nusatranslation_senti_mak	MAK	인도네시아	1200
nusatranslation_senti_min	MIN	인도네시아	1200
nusatranslation_senti_mui	MUI	인도네시아	500
nusatranslation_senti_rej	REJ	인도네시아	500
nusatranslation_senti_sun	SUN	인도네시아	1200
nusax_senti_ind	IND	인도네시아	400
nusax_senti_ace	ACE	인도네시아	400
nusax_senti_jav	JAV	인도네시아	400
nusax_senti_sun	SUN	인도네시아	400
nusax_senti_min	MIN	인도네시아	400
nusax_senti_bug	BUG	인도네시아	400
nusax_senti_bbc	BBC	인도네시아	400
nusax_senti_ban	BAN	인도네시아	400
nusax_senti_nij	NIJ	인도네시아	400
nusax_senti_mad	MAD	인도네시아	400
nusax_senti_bjn	BJN	인도네시아	400
nusax_senti_eng	ENG	비토착민	400
인도 영어	인도네시아어	인도네시아	1011

표 5: SEACrowd NLU 평가에 사용된 감정 분석 데이터 하위 집합.

하위 집합 ID	언어	지역	샘플 수
NLI → *_seacrowd_pairs			
인도네시아어 NLI	IND		인도네시
	아	5183	
wrete	IND	인도네시아	100
snli_indo	IND	인도네시아	9823
myxnli	MYA	미얀마	5010
xnli.tha	THA	태국	5010
xnli.vie	VIE	베트남	5010

표 6: SEACrowd NLU 평가에 사용된 NLI 데이터 하위 집합.

시스, 감정 분류, 법률 분류 및 기타. 이 데이터는 (id, text, label)로 구성되며, id는 데이터 세트의 고유 행 식별자를, text는 입력 텍스트를, label은 결정론적 목표 변수를 나타냅니다.

- **다중 레이블 텍스트 분류(TEXT MULTI).** 이 스키마는 증오 발언 탐지 및 측면 기반 감정 분석에 사용될 수 있습니다. (id, text, labels)로 구성되며, id는 데이터 세트의 고유 행 식별자를, text는 입력 텍스트를, labels는 결정론적 목표 변수 목록을 나타냅니다.
- **텍스트-투-텍스트(T2T)**. 이 스키마는 기계 번역, 요약 및 의역에 사용될 수 있습니다. (id, text_1, text_2, text_1_name, text_2_name)으로 구성되며, 여기서 id는 데이터 세트의 고유 행 식별자를 나타내고, text_1과 text_2는

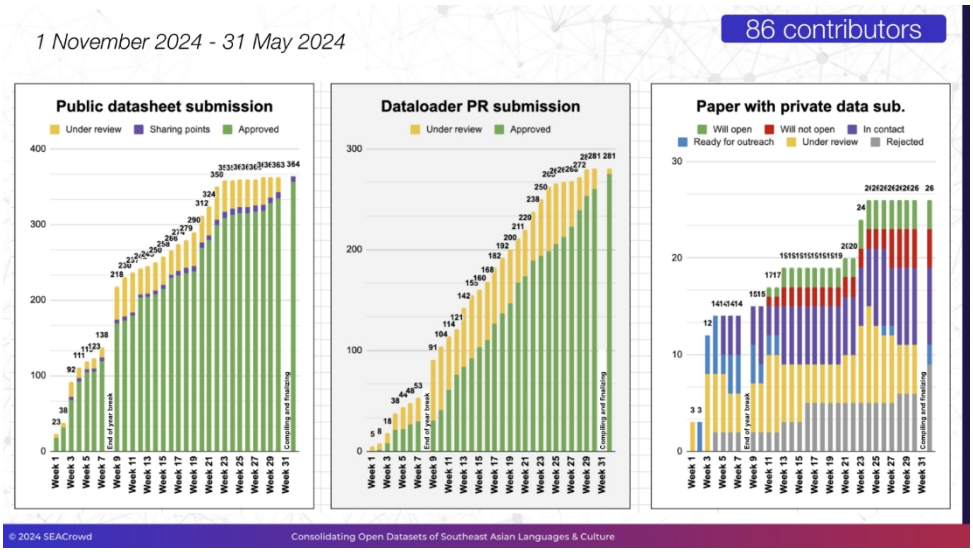


그림 9: SEACrowd의 누적 제출 건수 주간 현황 업데이트.

입력 텍스트 쌍을 나타내며, text_1_name과 text_2_name은 입력 텍스트 쌍의 이름을 나타냅니다(예: 번역 입력 텍스트 쌍의 경우 ind와 jav, 요약 입력 텍스트 쌍의 경우 document와 summary).

- **시퀀스 라벨링(SEQ LABEL)**. 이 스키마는 명명된 엔티티 인식(NER), 품사 태깅 등에 사용될 수 있습니다. (id, tokens, labels)로 구성되며, id는 데이터셋의 고유 행 식별자를, tokens는 입력 텍스트의 토큰 목록을, labels는 토큰에 대한 목표 목록을 나타냅니다.
- **질의응답(QA)**. 이 스키마는 추출형 QA, 객관식 QA 등에 활용될 수 있습니다. (id, question_id, document_id, question, type, choices, context, answer)로 구성되며, id는 데이터셋의 고유 행 식별자를, question_id는 질의의 고유 식별자를, document_id는 문맥 문서의 고유 식별자를, question은 답변할 입력 질의를, type은 QA 작업 유형(예: 추출형, 객관식, 개방형 생성, 폐쇄형 생성 등)을, choices는 답변 선택지 목록(필요한 경우)을, context는 질문의 배경 정보 역할을 하는 문장(필요한 경우)을, answer는 질문에 대한 정답(필요한 경우)을 나타냅니다.
- **단일 레이블 텍스트 쌍 분류(PAIRS).** 이는 텍스트 함축 관계 및 다음 문장 예측에 활용될 수 있습니다. 이는 (id, text_1, text_2, label)로 구성되며, id는 데이터셋의 고유 행 식별자를, text_1과 text_2는 입력 텍스트 쌍을, label은 목표 변수를 나타냅니다.
- **연속값 또는 회귀를 이용한 단일 레이블 텍스트 쌍 분류(PAIRS SCORE).** 이는 답안 채점 및 의미적 텍스트 유사도 계산에 활용될 수 있습니다. (id, text_1, text_2, label)로 구성되며, id는 데이터셋의 고유 행 식별자를, text_1과 text_2는 입력 텍스트 쌍을, label은 연속값 형태의 목표 변수를 나타냅니다.
- **다중 레이블 텍스트 쌍 분류(PAIRS MULTI)**. 형태론적 굴절 분석에 활용 가능합니다. (id, text_1, text_2, labels)로 구성되며, id는 데이터셋의 고유 행 식별자를, text_1과 text_2는 입력 텍스트 쌍을, labels는 목표 변수 목록을 나타냅니다.
- **지식 기반(KB)**. 이 스키마는 구성 요소 구문 분석, 종속성 구문 분석, 동일참조 해결, 대화 시스템 및 기타 복잡한 구조를 가진 작업에 사용될 수 있습니다. (id, passages, entities, events, coreferences, relations)로 구성됩니다. 복잡한 구조를 고려하여 독자들은 지식 기반 스키마의 구현을 살펴보길 권장합니다.
- **트리(TREE)**. 이 스키마는 구성성 분석에 사용될 수 있으며, 하위 노드 요소와 트리 계층 구조를 가진 문서를 가정합니다. (id, passage,

하위 집합 ID	언어	지역	샘플 수
주제 분류 → *_seacrowd_text			
gklmip_newsclass	KHM	캄보디아	1436
인도네시아 뉴스 데이터 세트	IND	인도네시아	2627
uit_vion	VIE	베트남	26000
sib_200_ace_Arab	ACE	인도네시아	204
sib_200_ace_Latn	ACE	인도네시아	204
sib_200_ban_Latn	BAN	인도네시아	204
sib_200_bjn_Arab	BJN	인도네시아	204
sib_200_bjn_Latn	BJN	인도네시아	204
sib_200_bug_Latn	BUG	인도네시아	204
sib_200_ceb_Latn	CEB	필리핀	204
sib_200_ilo_Latn	ILO	필리핀	204
sib_200_ind_Latn	IND	인도네시아	204
sib_200_jav_Latn	JAV	인도네시아	204
sib_200_kac_Latn	KAC	미얀마	204
sib_200_khm_크메르	KHM	캄보디아	204
sib_200_lao_Laoo	LAO	라오스	204
sib_200_lus_Latn	LUS	미얀마	204
sib_200_min_Arab	MIN	인도네시아	204
sib_200_min_Latn	최소	인도네시아	204
sib_200_mya_Mymr	MYA	미얀마	204
sib_200_pag_Latn	PAG	필리핀	204
sib_200_shn_Mymr	SHN	미얀마	204
sib_200_sun_Latn	SUN	인도네시아	204
sib_200_tgl_Latn	필리핀	필리핀	204
sib_200_tha_태국어	THA	태국	204
sib_200_vie_Latn	VIE	비토착민	204
sib_200_war_Latn	전쟁	필리핀	204
sib_200_zsm_Latn	ZSM	말레이시아	204
nusaparagraph_topic_btk	BTX	인도네시아	500
nusaparagraph_topic_bew	BEW	인도네시아	800
nusaparagraph_topic_bug	BUG	인도네시아	300
nusaparagraph_topic_jav	JAV	인도네시아	800
nusaparagraph_topic_mad	MAD	인도네시아	700
nusaparagraph_topic_mak	MAK	인도네시아	700
nusaparagraph_topic_min	최소	인도네시아	800
nusaparagraph_topic_mui	MUI	인도네시아	400
nusaparagraph_topic_rej	REJ	인도네시아	350
nusaparagraph_topic_sun	SUN	인도네시아	900

표 7: SEACrowd NLU 평가에 사용된 주제 분류 데이터 하위 집합.

하위 집합 ID	언어	지역	샘플 수
상식적 추론 → *_seacrowd_text/qa			
emotes_3k_tgl	FIL	필리핀	2905
emotes_3k_eng	ENG	비토착민	2905
인도 스토리 클로즈	인도네시아어	인도네시아	1135
xstorycloze_id	IND	인도네시아	1511
xstorycloze_my	MYA	미얀마	1511

표 8: SEACrowd NLU 평가에 사용된 상식 추론 데이터 하위 집합.

노드), 여기서 id는 데이터 세트의 고유 행 식별자를 나타내고, 패시지는 해당 특정 id에 해당하는 패시지를 나타냅니다. 이 패시지는 (id, type, text, offsets)로 구성됩니다. 노드는 해당 특정 id에 해당하는 노드를 나타내며, 이 노드는 (id, type, text, offsets, subnodes)로 구성됩니다.

- **대화형 채팅(CHAT)**. 이 스키마는 대화형 채팅 및/또는 다중 회화(multi-turn conversation)에 사용될 수 있습니다. (id, input, output, meta)로 구성되며, 여기서 id는 데이터 세트의 고유 행 식별자를 나타내고, input은 콘텐츠로 구성된 시퀀스를 나타냅니다.

부분집합 ID	언어	지역	샘플 수
표준 테스트 QA → *_seacrowd_qa			
indommlu_ind	IND	인도네시아	14979
indommlu_ban	BAN	인도네시아	14979
indommlu_mad	MAD	인도네시아	14979
indommlu_mak	MAK	인도네시아	14979
indommlu_sun	SUN	인도네시아	14979
indommlu_jav	JAV	인도네시아	14979
indommlu_bjn	BJN	인도네시아	14979
indommlu_abl	ABL	인도네시아	14979
indommlu_nij	NIJ	인도네시아	14979
seaeval_cross_mmlu_ind	IND	인도네시아	150
seaeval_cross_mmlu_vie	VIE	베트남	150
seaeval_cross_mmlu_zlm	ZSM	말레이시아	150
seaeval_cross_mmlu_fil	필리핀	필리핀	150
seaeval_cross_logiqa_ind	IND	인도네시아	176
seaeval_cross_logiqa_vie	VIE	베트남	176
seaeval_cross_logiqa_zlm	ZSM	말레이시아	176
seaeval_cross_logiqa_fil	필리핀	필리핀	176
m3exam_jav	JAV	인도네시아	371
m3exam_tha	THA	태국	2168
m3exam_vie	VIE	베트남	1789
okapi_m_arc_ind	IND	인도네시아	1170
okapi_m_arc_vie	VIE	베트남	1170
문화적 QA → *_seacrowd_qa			
copal_colloquial	IND	인도네시아	559
xcopa_tha	THA	태국	500
xcopa_vie	VIE	베트남	500
xcopa_ind	IND	인도네시아	500
seaeval_sg_eval_eng	ENG	비토착종	103
seaeval_ph_eval_eng	ENG	비토착종	100
mabl_ind	IND	인도네시아	1140
mabl_jav	JAV	인도네시아	600
mabl_sun	SUN	인도네시아	600
독해 QA → *_seacrowd_qa			
belebele_ceb_latn	CEB	필리핀	900
belebele_ilo_latn	ILO	필리핀	900
belebele_ind_latn	IND	인도네시아	900
belebele_jav_latn	JAV	인도네시아	900
belebele_kac_latn	KAC	미얀마	900
belebele_khm_khmr	KHM	캄보디아	900
belebele_lao_laoo	라오스	라오스	900
belebele_mya_mymr	MYA	미얀마	900
belebele_shn_mymr	SHN	미얀마	900
belebele_sun_latn	SUN	인도네시아	900
belebele_tgl_latn	필리핀	필리핀	900
belebele_tha_thai	THA	태국	900
belebele_vie_latn	VIE	베트남	900
belebele_war_latn	WAR	필리핀	900
belebele_zsm_latn	ZSM	말레이시아	900

표 9: SEACrowd NLU 평가에 사용된 객관식 QA 데이터 하위 집합.

입력 프롬프트 역할과 프롬프트를 입력하는 엔티티의 역할을 나타내며, 출력(output)은 해당 입력 프롬프트에 대한 답변을, 메타(meta)는 스키마의 유연성을 허용하기 위한 관련 세부 정보를 나타냅니다(필요한 경우).

- **종단간 작업 지향 대화(TOD).** 이 스키마는 종단간 작업 지향 대화에 사용될 수 있습니다. 이 스키마는 (dialogue_idx, dialogue)로 구성되며, dialogue_idx는 대화의 고유한 행 식별자를 나타내고, dialogue는 턴 레이블, 시스템 발화, 턴 idx, 신념 상태(슬롯 및 액트로 구성됨), 사용자 발화 및 시스템 액트와 같은 핵심 세부 정보를 나타냅니다.

하위 집합 ID	언어		# 샘플
추출형 및 요약형 QA → *_seacrowd_qa			
facqa	IND	인도네시아	311
iapp_squad	THA	태국	739
qasina	IND	인도네시아	500
mkqa_khm	KHM	캄보디아	10000
mkqa_zsm	ZSM	말레이시아	10000
mkqa_tha	THA	태국	10000
mkqa_vie	VIE	베트남	10000

표 10: SEACrowd NLG 평가에 사용된 추출형 및 추상형 QA 하위 집합.

하위 집합 ID	언어	지역	샘플 수
요약 → *_seacrowd_t2t			
lr_sum_ind	IND	인도네시아	500
lr_sum_vie	VIE	베트남	1460
lr_sum_lao	라오스	라오스	1496
lr_sum_tha	THA	태국	500
lr_sum_khm	KHM	캄보디아	486
lr_sum_mya	MYA	미얀마	990
xl_sum_mya	MYA	미얀마	570
xl_sum_ind	IND	인도네시아	4780
xl_sum_tha	THA	태국	826
xl_sum_vie	VIE	베트남	4013

표 11: SEACrowd NLG 평가에 사용된 요약 데이터 하위 집합.

하위 집합 ID	언어	지역	샘플 수
이미지 캡셔닝 → *_seacrowd_imtext			
xm3600_fil	FIL	필리핀	2760
xm3600_id	IND	인도네시아	2775
xm3600_th	THA	태국	2798
xm3600_vi	VIE	베트남	2855

표 12: SEACrowd VL 평가에 사용된 이미지 캡션 데이터 하위 집합.

F2 음성

- **음성-텍스트(SPTEXT).** 이는 음성 인식, 텍스트 음성 변환(TTS) 또는 음성 합성, 음성-텍스트 변환에 사용될 수 있습니다. 이것은 (id, path, audio, text, speaker_id, metadata)로 구성되며, id는 데이터 세트의 고유한 행 식별자를 나타내고, path는 입력 오디오 소스의 파일 경로를 나타내며, audio는 해당 경로에서 로드된 오디오 데이터를 나타내고, text는 입력 텍스트를 나타내며, speaker_id는 화자의 고유 식별자를 나타내고, metadata는 화자의 연령 및 성별과 같은 관련 세부 정보(필요한 경우)를 나타냅니다.
- **음성-음성(S2S)**. 음성-음성 번역에 활용될 수 있습니다. 이 데이터는 (id, path_1, audio_1, text_1, metadata_1, path_2, audio_2, text_2, metadata_2)로 구성되며, 여기서 id는 데이터셋의 고유 행 식별자를 나타내고, path_1과 path_2는 각각의 입력 오디오 소스 파일 경로를 나타내며, audio_1과 audio_2는 해당 경로에서 로드된 오디오 데이터를, text_1과 text_2는 입력 텍스트를, metadata_1과 metadata_2는 화자의 연령 및 성별(필요한 경우)과 같은 관련 세부 정보를 나타냅니다.
- **음성 분류(SPEECH)**. 이 스키마는 단일 레이블 용도로만 음성 분류, 음성 언어 식별 및 음성 감정 인식에 사용될 수 있습니다. 이 스키마는 (id, path, audio, speaker_id, labels, metadata)로 구성되며, id는 데이터셋의 고유 행 식별자를, path는 입력 오디오 소스의 파일 경로를, audio는 해당 경로에서 로드된 오디오 데이터를, speaker_id는 화자의 고유 식별자를, labels는 해당 특정 음성의 레이블(단일 레이블만 가능)을, metadata는 화자의 연령 및 성별과 같은 관련 세부 정보를 나타냅니다 (필요한 경우).
- **다중 레이블 음성 분류(SPEECH MULTILABEL)**. 이 스키마는 다중 레이블 용도로만 음성 분류, 음성 언어 식별 및 음성 감정 인식에 사용할 수 있습니다. 이 스키마는 (id, path, audio, speaker_id, labels, metadata)로 구성되며, id는 데이터 세트의 고유한 행 식별자를 나타내고, path는 입력 오디오 소스의 파일 경로를 나타내며, audio는 해당 경로에서 로드된 오디오 데이터를 나타내고, speaker_id는 화자의 고유 식별자를 나타내며, labels는 특정 음성의 라벨 시퀀스를 나타내며(다중 라벨만 가능), 메타데이터는 화자의 연령 및 성별(필요한 경우)과 같은 관련 세부 정보를 나타냅니다.

F3 VL

- **이미지-텍스트(IMTEXT)**. 이 스키마는 이미지 캡션 생성, 텍스트-이미지 생성, 비전-언어 사전 훈련에 사용될 수 있습니다. (id, text, image_paths, metadata)로 구성되며, id는 데이터셋의 고유 행 식별자를, text는 입력 텍스트를, image_paths는 입력 이미지 소스 경로 목록을, metadata는 시각적 개념 및 라벨(필요한 경우)과 같은 관련 세부 정보를 나타냅니다.
- **일반 이미지 분류(IMAGE)**. 이 스키마는 이미지 분류에 사용될 수 있습니다.

하위 집합 ID	언어	지역	샘플 수
ASR → *_seacrowd_	sptext		
asr_ibsc	IBA	브루나이	473
commonvoice_120_ind	IND	인도네시아	3647
commonvoice_120_tha	THA	태국	10964
commonvoice_120_cnh	CNH	미얀마	763
commonvoice_120_vie	VIE	베트남	1302
fleurs_ind	IND	인도네시아	687
fleurs_jav	JAV	인도네시아	728
fleurs_tha	THA	태국	1021
fleurs_lao	라오스	라오스	405
fleurs_mya	MYA	미얀마	880
플뢰르_크멤	KHM	캄보디아	771
fleurs_vie	VIE	베트남	857
fleurs_zlm	ZLM	말레이시아	749
fleurs_fil	FIL	필리핀	964
fleurs_ceb	CEB	필리핀	541
indspeech_newstra_ethnicsr_nooverlap_jav	JAV	인도네시아	1000
indspeech_newstra_ethnicsr_nooverlap_sun	SUN	인도네시아	1000
indspeech_newstra_ethnicsr_nooverlap_ban	BAN	인도네시아	1000
indspeech_newstra_ethnicsr_nooverlap_btk	BTX	인도네시아	1000

표 13: SEACrowd 음성 평가에 사용된 ASR 데이터 하위 집합.

단일 레이블 및 다중 레이블 모두를 포함합니다. 이 데이터는 (id, labels, image_path, metadata)로 구성되며, id는 데이터 세트의 고유한 행 식별자를, labels는 특정 이미지의 레이블을(단일 레이블 및 다중 레이블 가능), image_path는 입력 이미지 소스의 경로 목록을, metadata는 시각적 개념 및 레이블(필요한 경우)과 같은 관련 세부 정보를 나타냅니다.

- **이미지 질문 답변(IMQA)**. 이 스키마는 이미지/시각적 질문 답변에 사용될 수 있습니다. 이 스키마는 (id, question_id, document_id, questions, type, choices, context, answer, image_paths, meta)로 구성됩니다. 여기서 id는 데이터셋의 고유 행 식별자를, question_id는 질문의 고유 식별자를, document_id는 문맥 문서의 고유 식별자를 나타냅니다. question은 답변할 입력 질문을, type은 QA 작업 유형(예: 추출형, 객관식, 개방형 생성, 폐쇄형 생성, 등), choices는 답변 선택지 목록(필요한 경우), context는 질문의 배경 정보로 사용되는 문장(필요한 경우), answer는 질문에 대한 정답(필요한 경우), image_path는 입력 이미지 소스 경로 목록, metadata는 스키마의 유연성을 허용하기 위한 관련 세부 정보(필요한 경우)를 나타냅니다.
- **일반 영상-텍스트 변환(VIDEO)**. 이 스키마는 영상-텍스트 검색 및 영상 자막 생성에 사용될 수 있습니다. (id, video_path, text, metadata)로 구성되며, id는 denotes 데이터 세트의 고유 행 식별자를 나타내며, video_path는 입력 비디오 소스의 파일 경로를 나타내고, text는 해당 특정 프레임/비디오와 연관된 텍스트를 나타내며, metadata는 비디오의 해상도, 지속 시간, FPS(필요한 경우)와 같은 관련 세부 정보를 나타냅니다.

G SEA 평가를 위한 추가 세부사항

G.1 데이터셋

표 5, 6, 7, 8, 9는 NLU 평가에 사용된 데이터 하위 집합의 세부 사항을 제공합니다. 감정 분석 데이터셋은 원래 NusaX(Winata et al., 2023), NusaTranslation(Cahyawijaya et al., 2023b), SentiTaglish[20] , SmSA (Purwari-anti and Crisdayanti, 2019), PRDECT-ID (Sutoyo et al., 2022), 인도네시아어-영어 혼용 감정 데이터 (Astuti et al., 2023), 카로어 트윗 감정 데이터 (Karo et al., 2022), 태풍 요란다 감정 분석(Imperial et al., 2019), GKLMIP 크메르 감정 분석(Jiang et al., 2022), Wisesight 감정 분석 코퍼스([21]) ,필리핀-타갈로그 제품 리뷰 감정 분석([22]) ,그리고 인도네시아 모바일 앱 리뷰의 다중 레이블 감정 분석(Riccosan and Saputra, 2023). 주제 분류 데이터셋은 원래 NusaParagraph (Cahyawijaya et al., 2023b), UIT-ViON (Tran et al., 2021), SIB-200 (Adelani et al., 2024), GKLMIP 크메르어 뉴스(Jiang et al., 2022), 인도네시아어 뉴스(Muzad and Rahutomo, 2016)에서 가져왔습니다. 자연어 추론 데이터셋은 원래 IndoNLI (Mahendra et al., 2021), WreTe (Setya and Mahendra, 2018), SNLI Indo (Putra et al., 2024), MyXNLI([23])및 XNLI (Conneau et al., 2018)에서 가져온 것입니다. 상식 이해 soning 데이터셋은 원래 XStoryCloze(Lin et al., 2022), IndoCloze(Koto et al., 202 et al., 2022), IndoCloze (Koto et al., 2022), EMoTES-3K (Catapang and Visperas, 2023)에서 가져왔습니다.

오픈 도메인 QA 데이터 세트는 원래 In-doMMLU (Koto 외, 2023b), SeaEval (Wang 외, 2023), M3Exam (Zhang 외, 2023b) 및 Okapi (Dac Lai 외, 2023)에서 가져온 것입니다. 문화적 QA 데이터셋은 원래 COPAL-ID (Wibowo et al., 2023), XCOPA (Ponti et al., 2020), SeaEval (Wang et al., 2023), Multilingual Fig-QA (Kabra et al.,

[20] https://huggingface.co/datasets/ccosme/SentiTaglishProductsAndServices

[21] https://github.com/PyThaiNLP/wisesight-감정

[22] https://github.com/EricEchemane/Filipino-타갈로그어-제품-리뷰-감성-분석

[23] https://huggingface.co/datasets/akhtet/myXNLI

부분집합 ID		언어	지역	샘플 수
Eng → XX	XX → Eng			
*MT (Eng ⇔ XX) → *_seacrowd_t2t*				
리오_앤_센트럴_플로레스_영어_lji	lio_and_central_flores_lji_eng	LJL	인도네시아	1658
flores200_eng_Latn_ace_Latn	flores200_ace_Latn_eng_Latn	ACE	인도네시아	1012
flores200_eng_Latn_ban_Latn	flores200_ban_Latn_eng_Latn	BAN	인도네시아	1012
flores200_eng_Latn_bjn_Latn	flores200_bjn_Latn_eng_Latn	BJN	인도네시아	1012
flores200_eng_Latn_bug_Latn	flores200_bug_Latn_eng_Latn	BUG	인도네시아	1012
flores200_eng_Latn_ceb_Latn	flores200_ceb_Latn_eng_Latn	CEB	필리핀	1012
flores200_eng_Latn_ilo_Latn	flores200_ilo_Latn_eng_Latn	ILO	필리핀	1012
flores200_eng_Latn_ind_Latn	flores200_ind_Latn_eng_Latn	IND	인도네시아	1012
flores200_eng_Latn_jav_Latn	flores200_jav_Latn_eng_Latn	JAV	인도네시아	1012
flores200_eng_Latn_kac_Latn	flores200_kac_Latn_eng_Latn	KAC	미얀마	1012
flores200_eng_Latn_khm_크메르어	flores200_khm_크메르어_영어_라틴 문자	KHM	캄보디아	1012
flores200_eng_Latn_lao_Laoo	flores200_lao_Laoo_eng_Latn	라오스	라오스	1012
flores200_eng_Latn_lus_Latn	flores200_lus_Latn_eng_Latn	LUS	미얀마	1012
flores200_eng_Latn_min_Latn	flores200_min_Latn_eng_Latn	MIN	인도네시아	1012
flores200_eng_Latn_mya_Mymr	flores200_mya_Mymr_eng_Latn	MYA	미얀마	1012
flores200_eng_Latn_pag_Latn	flores200_pag_Latn_eng_Latn	PAG	필리핀	1012
flores200_eng_Latn_shn_Mymr	flores200_shn_Mymr_eng_Latn	SHN	미얀마	1012
flores200_eng_Latn_sun_Latn	flores200_sun_Latn_eng_Latn	SUN	인도네시아	1012
flores200_eng_Latn_tha_태국어	flores200_tha_태국어_eng_Latn	THA	태국	1012
flores200_eng_Latn_vie_Latn	flores200_vie_Latn_eng_Latn	VIE	베트남	1012
flores200_eng_Latn_war_Latn	flores200_war_Latn_eng_Latn	WAR	필리핀	1012
flores200_eng_Latn_zsm_Latn	flores200_zsm_Latn_eng_Latn	ZSM	말레이시아	1012
ntrex_128_eng-US_ind	ntrex_128_ind_eng-US	IND	인도네시아	1997
ntrex_128_eng-US_mya	ntrex_128_mya_eng-US	MYA	미얀마	1997
ntrex_128_eng-US_fil	ntrex_128_fil_eng-US	FIL	필리핀	1997
ntrex_128_eng-US_khm	ntrex_128_khm_eng-US	KHM	캄보디아	1997
ntrex_128_eng-US_lao	ntrex_128_lao_eng-US	라오스	라오스	1997
ntrex_128_eng-US_zlm	ntrex_128_zlm_eng-US	ZSM	말레이시아	1997
ntrex_128_eng-US_tha	ntrex_128_tha_eng-US	THA	태국	1997
ntrex_128_eng-US_vie	ntrex_128_vie_eng-US	VIE	베트남	1997
ntrex_128_eng-US_hmv	ntrex_128_hmv_eng-US	HMV	베트남	1997
nusax_mt_eng_ind	-	IND	인도네시아	400
nusax_mt_eng_ace	nusax_mt_ace_eng	ACE	인도네시아	400
nusax_mt_eng_jav	nusax_mt_jav_eng	JAV	인도네시아	400
nusax_mt_eng_sun	nusax_mt_sun_eng	SUN	인도네시아	400
누삭스_MT_ENG_MIN	nusax_mt_min_eng	MIN	인도네시아	400
nusax_mt_eng_bug	nusax_mt_bug_eng	BUG	인도네시아	400
nusax_mt_eng_bbc	nusax_mt_bbc_eng	BBC	인도네시아	400
nusax_mt_eng_ban	nusax_mt_ban_eng	BAN	인도네시아	400
nusax_mt_eng_nij	nusax_mt_nij_eng	NIJ	인도네시아	400
nusax_mt_eng_mad	nusax_mt_mad_eng	MAD	인도네시아	400
nusax_mt_eng_bjn	nusax_mt_bjn_eng	BJN	인도네시아	400

표 14: SEACrowd NLG 평가에 사용된 영어와 SEA 언어 데이터 하위 집합 간의 MT.

2023). 독해 데이터 세트는 원래 Belebele (Bandarkar et al., 2023)에서 가져온 것입니다.

표 10, 11, 14는 NLG 평가에 사용된 데이터 하위 집합의 세부 사항을 제공합니다. 요약 데이터셋은 원래 LR-Sum(Palen-Michel and Lignos, 2023)과 XL-Sum(Hasan et al., 2021)에서 비롯되었습니다. 기계 번역 데이터셋은 원래 Lio 및 Central Flores 코퍼스(Elias, 2018), Flores-200(Costa-jussà et al., 2024), NTREX-128(Federmann et al., 2022)에서 가져왔습니다. 질의응답 데이터셋은 원래 FacQA (Purwarianti 외, 2007), QASiNa (Rizqul-lah 외, 2023), MKQA (Longpre 외, 2021) 및 Open Thai Wikipedia QA 데이터셋([24])에서 가져온 것입니다.

표 12와 13은 VL 및 음성 평가에 사용된 데이터 하위 집합의 세부 정보를 제공합니다.

[24] https://zenodo.org/records/4539916

이미지 캡셔닝 데이터셋은 원래 XM3600(Thapliyal et al., 2022)에서 비롯되었습니다. 음성 인식 데이터 세트는 원래 INDspeech NEW-STRA Ethnic 컬렉션(Sani 외, 2012), ASR Iban(Juan 외, 2015), FLEURS(Conneau 외, 2022) 및 Common Voice(Ardila 외, 2020)에서 가져온 것입니다.

G.2 기준선

표 20, 21, 22는 SEACrowd 평가(§3)에 사용된 기준 모델의 세부 사항을 보고합니다. 각 기준 모델에 대해 모델 크기, 원본 기반 모델, 훈련 코퍼스에서 사용된 언어, 모델 다운로드 URL에 대한 정보를 제공합니다. 원칙적으로 본 연구는 인터넷에서 구할 수 있는 모든 SEA 훈련 LLM을 획득하고 적합시키는 것을 목표로 하지 않습니다. 이는 계산 비용이 너무 크기 때문입니다. 오히려 우리는

Model	$\tau = 0.01$	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.7$	$\tau = 1.0$
상업적					
GPT-4	0.199	0.192	0.155	0.118	**0.066**
CommandR	0.201	0.198	0.185	0.168	0.126
영어					
미스트랄	0.161	0.160	0.159	0.162	0.150
Llama3	0.138	0.137	0.131	0.129	0.113
Falcon	0.274	0.272	0.238	0.250	0.211
다국어					
mT0	0.151	0.148	0.131	0.112	0.074
BLOOMZ	0.238	0.236	0.228	0.217	0.167
박트리아노스-라마	0.163	0.162	0.163	0.168	0.149
AYA-23	0.183	0.182	0.183	0.179	0.135
AYA-101	**0.112**	**0.109**	**0.095**	**0.085**	0.069
SEA 지역					
SEA-LION	0.250	0.242	0.204	0.164	0.102
SeaLLM v2.5	0.137	0.133	0.116	0.097	0.069
세일러	0.152	0.151	0.145	0.139	0.113
SEA 국가					
Cendol-mT5	0.407	0.404	0.378	0.328	0.200
Cendol-Llama2	0.294	0.290	0.267	0.232	0.149
메라크 v4	0.209	0.207	0.199	0.190	0.155
왕찬X-라마3	0.163	0.161	0.153	0.150	0.131
말레이시아 Llama3	0.181	0.181	0.179	0.176	0.143

표 15: 다양한 τ 값에 따른 인구 가중 지니 계수를 기반으로 한 기준선 간 언어 형평성. 지니 계수가 낮을수록 형평성이 높음.

모델	하이퍼파라미터	값
로지스틱 회귀	max_iter C	100 np.linspace(0.001, 10, 100)
순진 베이즈	알파 분포	np.linspace(0.001, 1, 50) 다항 뉴턴-베이지안
SVM	C 커널	1 ["rbf", "linear"]

표 16: 그리드 검색을 통한 Translationese 예측을 위한 고전적 모델의 하이퍼파라미터.

공개적으로 이용 가능한 선별된 모델들의 탐색을 시작하여, SEACrowd를 통해 집계된 NLU, NLG, 음성, 비전 작업에 대한 벤치마킹을 통해 SEA 언어에 대한 기초 능력 평가의 기준선으로 활용하기 위함이다.

표에 나열된 다양한 모델을 검토하는 과정에서, 우리는 규모, 개방성, 동남아시아 언어(SEA) 커버리지 측면에서 모델 변이의 다양성을 우선시했습니다. NLP 작업에서는 주요 실험을 위해 5가지 LLM 그룹을 다루었습니다: 영어 전용, 다국어, 지역별, 국가별 모델입니다. 지시문 튜닝된 LLM은 미지 작업에 대한 일반화 능력을 보여줍니다(Wei et al., 2021; Sanh et al., 2021; Ouyang et al., 2022). 이들 LLM 중 일부는 다국어 기반 모델을 토대로 하여 언어 간 일반화 능력이 뛰어납니다(Muennighoff et al., 2022; Adilazuarda et al., 2023; Zhang et al., 2023a). NLU의 경우 가중 F1 점수를 계산하며, 오픈소스 기준선에서는 로그 가능도, 상용 기준선에서는 문자열 매칭을 통해 답을 얻습니다.

음성 벤치마크의 경우, 단 두 가지 모델 패밀리만...

모델	3-라벨	HT 대 MT-Nat MT 대 HT-Nat Nat 대 HT-MT		
LR (TF-IDF)	39.73	53.03	56.01	75.20
LR (BoW)	45.63	55.90	61.39	75.60
NB (TF-IDF)	33.43	49.53	50.55	73.05
NB (BoW)	33.70	49.10	50.64	71.26
SVM (TF-IDF)	39.55	52.63	55.10	76.40
SVM (BoW)	46.84	56.85	61.40	75.65
mDeBERTa	51.51	64.77	59.16	**79.08**

표 17: 언어별로 평균한 번역어 분류기 결과(정확도).

국가	소속	원산지
인도네시아	16	31
말레이시아	0	1
필리핀	3	7
싱가포르	13	2
태국	1	2
베트남	0	1
호주	1	0
브라질/스웨덴	0	1
캐나다	1	0
중국	2	8
이집트	0	1
독일	0	2
홍콩	2	0
인도	0	1
아일랜드	1	0
일본	3	0
네덜란드	0	1
아랍에미리트	5	0
영국	4	0
미국	9	1
우즈베키스탄	0	2

표 18: 소속 국가 및 출신 국가별 저자 인구 통계.

다국어 모델과 특정 동남아시아 언어에 미세 조정된 모델이 제공됩니다. 비전 작업의 경우 영어 전용 모델과 다국어 모델 하나를 다루었습니다. 이러한 모델들은 이미지-텍스트 정렬에 대해 사전 훈련된 시각적 백본(예: CLIP (Radford et al., 2021))을 활용하여 이미지 특징을 기존 사전 훈련된 언어 모델의 입력 공간으로 투영합니다. 요약하면, 우리는 주로 HuggingFace에서 쉽게 접근할 수 있는 오픈 모델을 탐구했지만, 성능 벤치마킹, 재현성 및 향후 연구를 위한 확장을 위해 GPT-4 및 Whisper V3와 같은 상용 모델도 포함했습니다.

G.3 프롬프트

표 23, 24, 25는 NLU, NLG, VL 평가(§3)에 사용된 필기 프롬프트 템플릿을 설명합니다. 모든 작업에 대해 제로샷 프롬프팅 절차를 기준 설정으로 사용했습니다. 작업 복잡성과 자원 보유 자원봉사자들의 작업량 분배로 인해, 일부 설정에서는 목표 공개 일정에 맞춰 결과를 확보하기 위해 실험 절차를 제한했습니다.

NLU의 경우 상식 추론, 질문응답, NLI 등 핵심 작업별 데이터셋마다 세 가지 프롬프트 스타일을 탐구했습니다. NLG 및 VL과 같이 더 많은 컴퓨팅 자원이 필요한 고난도 작업의 경우 단일 통일된 프롬프트 스타일만 사용했으나, VL의 경우 SEA 언어(필리핀어, 인도네시아어, 태국어, 베트남어)로 번역된 프롬프트도 함께 탐구했습니다.

G.4 평가 결과

표 26과 27은 언어별 NLU 및 NLG 결과를 설명합니다.

G.5 언어 평등성 결과

표 15는 지니 계수 계산에서 언어 사용자 수의 가중치를 달리하여 평가에 사용된 대규모 언어 모델(LLMs)의 언어 평등도를 제시합니다.

H 번역어 분류기 보조 세부사항

H.1 훈련 및 평가 데이터

번역체 분류기 훈련 및 평가를 위한 각 데이터 하위 집합의 텍스트 수집 방법을 표 28과 29에서 각각 수동으로 선정하고 검증합니다. 이 검증은 관련 논문, 도메인 및 주석 방법을 확인하여 수행됩니다. 데이터 하위 집합의 텍스트가 기계 번역 또는 인간 번역의 산물인 경우, 이를 번역체로 간주합니다. 인간이 생성한 텍스트로 구성된 데이터 하위 집합은 자연 데이터로 라벨링합니다.

H.2 실험

우리는 기계 학습 모델이 인간 생성/자연 샘플(Nat), 인간 번역 샘플(HT), 기계 번역 샘플(MT)을 구분하는 능력을 평가하고자 합니다. 본 연구에서는 고전적 머신러닝 기법을 활용한 분류기 훈련과 학습 성능 향상을 위한 mDeBERTa 모델의 미세 조정을 수행합니다. 또한 두 레이블 클래스를 하나로 통합하여 실험함으로써 이들 레이블 간 구별의 예측 난이도를 평가합니다. 이 분석은 각 범주 간 샘플의 상대적 유사성에 대한 귀중한 통찰력을 제공합니다. 다음 섹션에서는 본 연구 방법론에 대한 포괄적인 개요를 제시합니다.

고전적 ML 우리는 세 가지 고전적 기계 학습 방법을 사용합니다: 1) 로지스틱 회귀(LR),
2) 나이브 베이즈(NB), 3) 서포트 벡터 머신(SVM)으로, 두 가지 다른 특징을 포함합니다. TF-IDF 및 단어백(BoW). 검증 세트에서 각 방법에 대한 최적 하이퍼파라미터를 찾기 위해 그리드 검색을 통한 하이퍼파라미터 튜닝을 수행하고, 테스트 세트 결과를 표 16에 보고합니다.

인코더 기반 LM 번역어 분류기 개발을 위해 인코더 전용 LM의 파인 튜닝을 탐구합니다. 백본 모델로 다국어 인코더 전용 LM인 mDeBERTa-v3$_{base}$모델[25](He et al., 2020, 2022)을 활용합니다. AdamW(Loshchilov and Hutter, 2019) 최적화기를 사용하여 학습률 1e-5, 배치 크기 256, 워밍업 단계 500으로 최대 10에포크 동안 모델을 훈련합니다. 검증 정확도에 기반한 3에포크 조기 종료(early stopping)를 적용합니다. 결과는 표 17에 제시합니다.

I SEA 언어 우선순위 지정 보충 세부사항

글로벌 유용성 지표(Blasi et al., 2022) 결과를 바탕으로, 수요(즉, 동남아시아 언어 사용자 수)와 현재 유용성(그림 10) 또는 자원 가용성(그림 11)을 기준으로 우선순위가 부여된 상위 20개 동남아시아 원주민 언어를 제시합니다.[26] 현재 유용성 평가에는 동남아시아 언어에서 최고 성능 모델 중 하나인 AYA-101의 성능 점수를 활용합니다. 현재 유용성(모델 능력)은 영어(ENG) 모델 성능 대비 상대적이며, 자원 가용성은 HuggingFace에서 한국어 데이터셋 약 500개 분량을 기준으로 상대적입니다. 한국어가 피벗 언어로 선택된 이유는 Joshi et al. (2020)에 의해 대부분의 언어보다 높은 자원 가용성을 가진 언어로 간주되기 때문입니다.

J 기여자 인구통계

표 18은 SEACrowd 내 저자들의 지리적 분포를 설명합니다.

K 연구 대상 언어

표 30-48은 SEACrowd가 다루는 동남아시아 원주민 언어 목록을 제시합니다. ISO 639-3 코드, 언어명, 지역 및 인구에 관한 정보는 (Eberhard 외, 2021; Hammarström 외, 2024; Project, 2024; Dryer 및 Haspelmath, 2013) 및 Wikipedia(27)에서 얻었습니다.

[25] https://huggingface.co/microsoft/mdeberta-v3-base
[26] https://github.com/SEACrowd/globalutility [27] https://www.wikipedia.org/

번호	이름	C. 점수
1	성스러운 러브니아	549
2	사무엘 카히아위자야	480
3	라흐마드 마헨드라	317
4	살사빌 마울라나 아크바르	243
5	레스터 제임스 V. 미란다	234
6	정신 용	164
7	제니퍼 산토소	164
8	엘리아나 아코	158
9	아크단 파딜라	157
10	조니벡 만수로프	132
11	파즈리 코토	121
12	조셉 마빈 임페리얼	118
13	장루오천	114
14	겐타 인드라 위나타	108
15	온노 P. 캄프만	107
16	조엘 루벤 앤토니 모니즈	93
17	무함마드 라비 술탄 하비비	92
18	프레데리쿠스 후디	83
19	세드릭 케	81
20	알함 피크리 아지	80
21	라일리 몬탈란	78
22	피랏 림콘초티왓	72
23	라이언 이그나티우스	56
24	조아니토 아길리 로포	50
25	윌리엄 닉슨	50
26	보예 F. 칼손	49
27	제임스 자야	48
28	리안디토 디안다루	48
29	유제 가오	48
30	William Tjhi	46
31	패트릭 아마데우스	46
32	빈 왕	44
33	얀 크리스티안 블레이즈 크루즈	43
34	첸시 화이트하우스	36
35	이반 할림 파르모낭간	36
36	마리아 켈리	36
37	세바스찬 루더	35
38	웬유 장	34
39	럭키 수산토	33
40	레이나드 아다 랴다	32
41	소니 라주아르디 헤르마완	30
42	댄 존 벨라스코	29
43	무함마드 데한 알 카우사르	29
44	윌리 피트라 헨드리아	29
45	야스민 모슬렘	29
46	노아 플린	28
47	무함마드 파리드 아딜라주아르다	27
48	하오천 리	27
49	Johanes Lee	27
50	R. 다만후리	27
51	슈오 쑨	27
52	무함마드 레자 코림	26
53	아미르베크 자니베코프	25
54	웨이 치 렁	25
55	Quyet V. Do	24
56	니클라스 윈니호프	24
57	탄라다 판수완	22
58	일함 피르다우시 푸트라	21
59	얀 쉬	21
60	아유 푸르와리안티	20
61	Ngee Chia Tai	20

표 19: 기여 점수 순으로 정렬된 공동 저자 목록.

L 공동 저자별 기여도 점수

표 19는 SEACrowd에서의 기여도 순으로 정렬된 공동 저자 목록을 제공합니다. 그들의 기여에 대한 자세한 내용은 기여도 추적에서 확인할 수 있습니다.

모델 이름	모델 크기	백본	본 뱅스	URL
			상업용	
GPT-4	해당 없음	GPT-4	해당 없음	https://openai.com/index/gpt-4/. NLU에는 turbo-2024-04-09를 를 NLU에, gpt-4o-2024-05-13을 NLG에 사용했습니다.
CommandR	36B	CommandR	2개의 동남아시아 언어(VIE, IND), 22개의 비동남아시아언 lang 어	https://cohere.comblogtcommandr
			영어	
미스트랄	7B	미스트랄	해당 없음	mistralai/미스트랄-7B-지침-v0.3
Llama3	8B	Llama3	해당 없음	meta-llama/Meta-Llama-3-8B-Instruct
Falcon	7B	Falcon	0 SEA 언어 (주로 영어)	tiiuae/falcon-7b-instruct
			다국어	
mT0	3B	mT5	2개의 동남아시아 언어(VIE, IND), 43개의 비동남아시아 언어	bigscience/mt0-xl
BLOOMZ	7B	BLOOM	2 SEA 언어 (VIE, IND), 43 비 SEA 언어	bigscience/bloomz-3b
박트리아X-라마	7B	라마	6개의 동남아시아 언어(인도네시아어, 베트남어, 크메르어, 미얀마어, 태국어, 캄보디아어, 베트남어), 46개의 비동남아시아 언어	MBZUAI/bactrian-x-llama-7b-merged
AYA-23	8B	명령	2개의 동남아시아 언어(인도어, 베트남어), 21개의 비동남아시아 언어	CohereForAI/aya-23-8B
AYA-101	13B	T5	9개의 동남아시아 언어(인도, 베트남, 태국, 스리랑카, 미얀마, 세르비아, 필리핀, 일본, 싱가포르), 92개의 비동남아시아 언어	CohereForAI/aya-101
			SEA 지역	
SEA-LION	7B	MPT	8개 동남아시아 언어(인도네시아어, 베트남어, 태국어, 태국어, 캄보디아어, 라오스어, 미얀마어), 3개 비동남아시아 언어	aisingapore/sea-lion-7b-instruct
SeaLLM v2.5	7B	SeaLLM	8개의 동남아시아 언어 (인도어, 베트남어, 태국어, 캄보디아어, 라오스어, 미얀마어)	SeaLLMs/SeaLLM-7B-v2.5
Sailor	7B	Qwen 1.5	5개의 동남아시아 언어(인도, 베트남, 라오스, 캄보디아, 태국), 2개의 비동남아시아 언어	sail/Sailor-7B-Chat
			동남아시아 국가	
Cendol-mT5	3B	mT5	1 SEA 언어 (인도네시아어), 18개의 인도네시아 현지 언어	indonlp/cendol-mt5-xl
Cendol-Llama2	7B	Llama2	1 SEA 언어(인도네시아어), 18개 지역 인도네시아어 변형	indonlp/cendol-llama2-7b
Merak v4	7B	Llama2	1 SEA 언어 (인도네시아어)	
왕잔X-라마3	8B	Llama3	4개의 동남아시아 언어(인도, 베트남, 태국, 미얀마) 및 26개의 비동남아시아 언어	airesearch/LLaMa3-8b-WangchanX-sft-Demo
말레이시아 Llama3	8B	Llama3	동남아시아 언어 1개 (ZLM)	mesolitica/malaysian-llama-3-8b-instruct-16k

표 20: SEACrowd NLU 및 NLG 평가에 사용된 LLM.

모델 이름	모델 크기	백본	사용 언어	URL
			다국어	
Whisper v3	15억 4천만	Whisper v3 89 비 SEA 및 9 SEA (IND, JAV, LAO, ZLM, MYA, TGL, THA, SUN, VIE)		openai/whisper-large-v3
MMS 1B	1B	MMS	993개 비-SEA 및 205개 SEA (ABP, ACE, ACN, AGN, AHK, AKB, ALJ, ALP, AMK, AOZ, ATB, ATQ, AYZ, BAN, BBC, BCL, BDG, BDQ, BEP, BGR, BHZ, BKD, BLT, BLX, BLZ, BNO, BPR, BPS, BRU, BTD, BTS, BTX, BVZ, BZI, CEB, CEK, CFM, CGC, CMR, CNH, CTD, DBJ, DNT, DNW, DTP, EIP, FRD, GBI, GOR, HAD, HAP, HIL, HLT, HNN, HVN, IBA, IFA, IFB, IFK, IFU, IFY, ILO, IND, ITV, JAV, JMD, KAC, KAK, KDT, KHG, KHM, KJE, KJG, KLW, KMD, KML, KNB, KNE, KPQ, KPS, KQE, KQR, KRJ, KRR, KVW, KXF, KXM, KYB, KYO, KYU, KZF, LAO, LAW, LBW, LCP, LEW, LEX, LHU, LIS, LJE, LJP, LLG, LND, LSI, MAD, MAK, MBB, MBT, MEJ, MHX, MHY, MIN, MKN, MNB, MNW, MNX, MOG, MQF, MQJ, MQN, MRW, MTD, MTJ, MVP, MWQ, MWV, MYA, MYL, NFA, NIA, NIJ, NLC, NLK, NOD, NPY, NST, OBO, PAG, PAM, PCE, PEZ, PLW, PMF, PPK, PRF, PRK, PRT, PSE, PTU, PWW, RAW, REJ, RGU, RHG, RIL, ROL, SAJ, SAS, SBL, SDA, SEA, SGB, SHN, SJM, SLU, SML, SNE, SUC, SUN, SXN, SYA, SZA, TBK, TBL, TBY, TCZ, TDJ, TES, TGL, THA, TIH, TLB, TNT, TOM, TVW, TWB, TWE, TWU, TXA, TXQ, UBL, URK, URY, VIE, WAR, WLO, XDY, XMM, XSB, XTE, YKA, YLI, YVA, ZLM, ZYP)	facebook/mms-1b-all
Seamless M4T v2	2.3B	Seamless	83개 비동남아시아 언어 및 9개 동남아시아 언어(인도네시아어, 일본어, 크메르어, 라오스어, 미얀마어, 태국어, 베트남어, 라오스어, 캄보디아어)	facebook/seamless-m4t-v2-large
			특정 언어에 대해 미세 조정됨	
XLSR 영어	300M	Wav2Vec2	46개 비동남아시아 언어 및 7개 동남아시아 언어(CEB, CNH, IND, LAO, TAM, TGL, VIE) 및 미세 조정 언어	jonatasgrosman/wav2vec2-large-xlsr-53-english
XLSR 인도네시아어-자비어-순다어				indonesian-nlp/wav2vec2-indonesian-javanese-sundanese
XLSR 인도네시아어 XLSR 태국어				Galuh/wav2vec2-large-xlsr-indonesian wannaphong/wav2vec2-large-xlsr-53-th-cv8-newmm
XLS-R 타갈로그어				sil-ai/wav2vec2-bloom-speech-tgl
XLS-R 버마어				sil-ai/wav2vec2-bloom-speech-mya
XLS-R 크메르어				vitouphy/wav2vec2-xls-r-300m-khmer
Whisper 인도네시아어 Whisper 태국어	15억 4천만	Whisper	89 비 동남아시아 및 9 동남아시아 (인도네시아, 일본, 라오스, 말레이시아, 미얀마, 태국, 싱가포르, 베트남)	cahya/whisper-large-id biodatlab/whisper-th-large-v3-combined
Whisper 크메르어				ksoky/whisper-large-khmer-asr

표 21: SEACrowd 음성 평가에 사용된 음성 모델.

모델 이름	모델 크기	백본	사전 훈련 이미지	URL
		영어		
LLaVA 1.5	해당 없음	해당 없음	해당 없음	해당 없음
LLaVA 1.6	7B	미스트랄-7B	해당 없음	liuhaotian/llava-v1.6-mistral-7b
Idefics2	8B	미스트랄-7B-v0.1	1.5B	HuggingFaceM4/idefics2-8b
PaliGemma	2B	Gemma-2B	해당 없음	google/paligemma-3b-pt-224
		다국어		
mBLIP	해당 없음	blip2-flan-t5-xl	해당 없음	Gregor/mblip-mt0-xl

표 22: SEACrowd VL 평가에 사용된 VLM.

번호	프롬프트 템플릿
	감정 분석
1	아래 텍스트의 감정을 분류하세요.\n[INPUT] => 감정 ([OPTIONS]): [LABEL_CHOICE]
2	다음 텍스트의 감정을 예측하세요.\n텍스트: [INPUT]\n[OPTIONS]로 답변하세요: [LABEL_CHOICE]
3	[INPUT]\n위의 텍스트의 감정은 무엇일까요? [OPTIONS]? [LABEL_CHOICE]
	주제 분류
1	아래 텍스트의 주제를 분류하세요.\n[INPUT] => 주제 ([OPTIONS]): [LABEL_CHOICE]
2	다음 텍스트의 주제를 예측하세요.\n텍스트: [INPUT]\n[OPTIONS]로 답변하세요: [LABEL_CHOICE]
3	[INPUT]\n위 텍스트의 주제는 무엇일까요? [OPTIONS]? [LABEL_CHOICE]
	상식 추론 → *_seacrowd_text
1	아래 텍스트의 도덕성을 분류하세요.\n[INPUT] => 도덕성 ([OPTIONS]): [LABEL_CHOICE]
2	다음 텍스트의 도덕성을 예측하세요.\n텍스트: [INPUT]\n[OPTIONS]로 답변하세요: [LABEL_CHOICE]
3	[INPUT]\n위의 텍스트의 도덕성은 무엇일까요? [OPTIONS]? [LABEL_CHOICE]
	상식 추론 → *_seacrowd_qa
1	질문: [질문]\n이 질문에 답하기 위해 더 합리적인 답변은 무엇인가요?\n선택지: [답변_선택지]\n정답: [라벨_선택지]
2	다음 질문: "[질문]"과 선택지: [답안 선택지]에 근거하여 정답은: [정답 선택지]입니다.
3	질문: [질문]\n선택지: [답안 선택지]\n주어진 질문에 대한 정답은: [정답 선택지]
	모든 QA
1	아래 글을 읽고 다음 질문에 답하십시오:\n글: [문맥]\n질문: [질문]\n선택지: [답안 선택지]\n정답: [정답 선택지]
2	[문맥]\n위의 글에 근거하여, [질문]\n선택지: [답안 선택지]\n정답: [정답 선택지]
3	[문맥]\n질문: [질문 내용]\n선택지:[답안 선택지]\n위 글을 참고하여 주어진 질문에 대한 정답은: [정답 선택지]
	NLI
1	가설: [INPUT_A]\n전제: [INPUT_B]\n질문: 가설과 전제 사이의 관계는 무엇인가? [OPTIONS]? [LABEL_CHOICE]
2	다음 전제와 가설이 주어졌습니다:\n가설: [INPUT_A]\n전제: [INPUT_B]\n논리적 관계(([OPTIONS]))를 결정하세요: [LABEL_CHOICE]
3	전제와 가설 사이에서 가장 적절한 관계([옵션])를 선택하세요:\n"[INPUT_B]"와 "[INPUT_A]" 사이의 관계: [라벨_선택]

표 23: NLU 작업에 사용된 프롬프트 템플릿.

번호	프롬프트 템플릿
	기계 번역(MT)
1	를 다음 텍스트 [SOURCE] 에서 [TARGET] 로 번역하십시오. 에 번역 을 직접 제공하십시오.\n텍스트: [INPUT]\n번역:
	요약
	추상적 및 추출적 QA
1	다음 텍스트를 요약하세요.\n텍스트: [INPUT]\n요약:
1	아래 글을 참고하여 다음 질문에 답하십시오:\n글: [문맥]\n질문: [질문]\n답변:

표 24: NLG 작업에 사용되는 프롬프트 템플릿.

언어	프롬프트 템플릿
	이미지 캡션 생성
ENG	다음 이미지를 [언어]로 캡션하세요.
FIL	다음 그림을 설명하세요.
IND	다음 그림을 설명하세요.

표 25: VL 평가에서 이미지 캡션 작업에 사용된 프롬프트 템플릿.

	ABL	ABS	ACE	BAN	BBC	BEW	BHP	BJN	BTX	BUG	CEB	ENG	FIL	ILO	IND	JAV	KAC	KHM	라오스	LUS	MAD	MAK	MIN	MUI	MYA	NIJ	PAG	REJ	SHN	SUN	THA	VIE	WAR	ZSM	전체
GPT-4	63.3	39.0	39.3	60.3	7.1	68.5	2.8	60.4	27.8	40.4	85.6	52.1	55.9	69.5	60.7	59.7	30.8	66.4	51.8	70.0	37.1	44.3	57.9	71.8	47.6	40.2	79.4	34.0	21.7	58.5	59.6	56.1	84.9	61.6	51.9
CommandR	50.1	80.8	57.6	62.8	47.4	81.8	58.2	57.1	57.3	57.9	66.7	69.4	51.1	56.8	58.3	61.2	36.5	41.5	33.8	63.9	61.9	58.4	66.4	81.7	34.8	53.3	75.6	69.6	35.4	63.2	42.7	55.9	67.6	55.7	58.0
미스트랄	36.7	53.6	46.4	49.6	33.0	59.3	44.3	44.6	44.3	48.8	53.5	69.2	48.4	49.1	52.5	46.7	33.2	29.8	30.7	56.1	45.7	44.8	51.2	62.6	27.4	40.1	69.2	48.6	31.9	48.3	40.8	45.2	54.4	49.6	46.8
라마3	37.3	40.3	43.2	48.9	34.8	44.5	32.6	42.2	38.5	42.9	51.2	59.5	45.2	46.7	49.2	44.4	28.5	34.6	30.3	46.8	39.0	38.0	43.6	49.2	35.2	39.6	60.5	38.5	31.1	45.2	43.8	45.5	50.3	49.0	42.6
팔콘	21.1	63.2	13.3	19.0	23.0	37.9	62.1	15.6	31.9	15.7	19.5	43.7	25.1	18.8	30.8	27.0	14.2	10.2	12.7	15.0	30.3	32.3	23.6	37.0	18.0	23.0	18.8	36.0	14.1	28.2	15.9	18.8	19.1	17.4	25.1
mT0	37.6	63.6	43.7	51.2	37.0	66.1	38.4	43.6	41.3	50.3	62.5	49.4	41.0	59.0	47.2	56.0	40.9	57.5	61.2	57.0	46.7	45.8	52.6	68.8	45.9	40.9	62.6	47.8	47.0	58.8	41.8	41.4	61.4	49.4	50.5
BLOOMZ	25.6	66.5	28.4	34.2	35.8	53.9	48.0	30.4	36.3	33.3	30.9	51.7	28.9	27.8	44.7	38.2	23.1	18.9	23.6	28.1	37.8	34.5	39.9	60.2	23.0	34.6	33.1	42.2	19.8	41.3	25.9	34.8	32.1	34.3	35.3
박트리아노스-라마	24.9	48.6	21.2	28.5	26.9	33.4	45.9	22.8	31.4	22.7	27.9	45.6	32.0	24.3	38.3	30.0	19.9	17.0	20.7	21.0	30.0	28.8	26.2	35.7	22.8	27.2	26.5	29.2	20.5	30.2	24.5	27.1	28.3	31.5	28.6
AYA-23	43.3	21.2	26.9	35.0	24.3	31.2	16.8	30.9	25.1	26.5	36.0	50.8	33.5	32.7	46.8	36.9	20.5	15.1	22.0	27.4	31.0	31.7	27.3	35.5	23.7	37.3	32.6	22.8	20.8	34.9	32.7	44.8	37.1	47.9	31.3
AYA-101	42.5	64.3	71.2	65.2	58.8	68.2	43.3	63.5	52.7	60.7	71.7	62.8	52.8	65.0	54.2	62.6	43.1	62.2	67.8	71.8	56.9	49.0	69.3	70.2	51.5	57.2	75.7	52.9	53.8	67.2	49.5	48.0	70.5	56.4	59.8
바다사자	10.3	62.3	13.5	16.5	21.3	35.3	60.3	13.4	31.8	15.2	13.6	26.6	20.6	10.2	27.6	21.4	8.7	16.8	15.2	12.5	26.8	28.3	22.8	34.6	23.0	16.0	14.4	34.1	9.7	23.4	16.3	14.7	14.2	13.3	21.9
SeaLLM v2.5	50.7	55.1	34.5	43.4	36.3	53.9	53.2	45.8	45.8	37.7	47.6	42.5	52.6	44.7	53.4	49.8	27.4	42.6	50.3	45.8	48.7	49.8	46.8	58.4	41.0	39.1	55.7	47.8	28.7	50.1	49.0	54.5	55.4	60.6	47.0
세일러	50.4	59.2	43.8	55.5	44.1	61.5	43.9	50.5	44.8	45.7	45.6	63.0	40.2	45.0	51.3	53.1	29.9	32.7	53.9	53.9	47.6	46.5	52.8	63.9	28.1	52.7	59.3	42.2	26.7	54.0	46.3	47.7	49.2	52.1	48.1
Cendol-mT5	15.0	98.5	38.3	42.3	84.7	99.4	95.6	33.3	92.6	68.6	14.1	38.7	23.8	12.2	33.4	50.5	10.4	20.3	15.3	9.6	76.5	70.2	65.2	99.6	16.6	52.6	12.8	98.9	7.2	56.6	26.4	14.7	15.1	15.9	44.8
Cendol-Llama2	17.5	80.0	30.8	33.5	60.6	49.3	73.4	27.9	45.1	32.3	18.7	36.8	21.4	17.8	37.4	35.1	14.7	13.2	15.9	15.0	46.3	38.1	37.1	51.6	19.9	40.3	17.7	47.7	16.5	38.5	20.6	17.3	18.5	18.4	32.5
메라크	37.0	68.6	37.7	48.3	36.4	66.1	60.1	41.4	50.4	47.8	42.4	59.6	37.9	39.7	48.5	48.4	27.9	24.2	28.0	44.3	51.7	51.0	50.5	70.3	27.2	40.0	58.6	57.9	28.6	50.8	29.3	35.3	43.7	47.1	45.2
왕잔X-라마3	38.4	59.3	26.8	35.2	35.0	43.3	56.9	31.6	38.3	31.2	32.3	57.6	36.6	29.3	45.0	38.7	23.7	24.3	25.1	26.6	40.4	41.4	34.8	43.6	31.6	37.0	31.2	42.9	23.5	39.8	36.5	38.4	31.3	37.0	36.6
말레이시아 라마3	38.9	62.3	38.1	41.9	39.2	46.9	58.3	39.5	40.5	35.9	37.8	55.5	34.5	33.1	48.6	42.6	24.7	18.9	20.4	33.6	42.1	41.0	42.5	48.5	22.2	39.6	46.8	41.1	19.6	44.0	33.7	34.6	37.7	49.9	39.2
전체	35.6	60.4	36.4	42.9	38.1	55.6	49.7	38.6	43.1	39.7	42.1	51.9	37.9	37.9	46.0	44.6	25.5	30.3	32.1	38.8	44.3	43.0	45.0	58.0	30.0	39.5	46.1	46.4	25.4	46.3	35.3	37.5	42.8	41.5	41.4

표 26: 언어별 가중 F1 점수 기준 NLU 평가 결과.

	ACE	BAN	BBC	BJN	BUG	CEB	FIL	HMV	ILO	IND	JAV	KAC	KHM	라오스	LJL	LUS	MAD	MIN	MYA	NIJ	PAG	SHN	SUN	THA	VIE	WAR	ZSM	전체
GPT-4	32.9	40.7	28.8	42.0	24.1	66.5	65.6	50.0	52.9	59.3	54.2	16.7	29.8	41.9	10.0	33.3	29.2	46.1	21.5	27.7	37.0	14.5	50.0	28.8	47.5	66.4	59.6	39.9
명령-R	19.6	26.1	16.4	30.0	16.0	44.3	52.5	16.8	29.4	57.9	32.6	8.8	8.7	14.2	6.0	19.5	17.2	31.6	9.5	18.4	20.4	8.9	27.5	24.3	46.8	34.4	50.1	25.5
미스트랄	12.4	15.0	10.0	13.9	11.1	28.5	37.2	10.2	15.9	28.6	15.4	7.3	8.7	10.8	4.2	11.7	9.5	18.0	5.7	12.4	17.5	9.5	14.8	15.1	25.1	22.4	31.1	15.6
Llama3	11.0	12.3	8.1	13.8	7.6	25.1	33.2	7.6	18.4	21.9	17.0	4.8	6.5	5.8	3.2	9.6	8.5	16.4	4.5	9.5	11.8	6.3	15.1	9.6	21.7	20.5	25.2	13.2
팔콘	7.3	9.5	8.2	8.3	7.9	18.6	23.6	6.6	9.7	15.3	7.7	6.0	3.1	3.1	4.2	9.3	6.6	11.8	1.8	8.7	12.9	4.5	7.7	2.4	13.5	13.5	17.0	9.2
mT0	4.8	5.6	3.7	5.7	3.1	4.6	6.8	4.5	3.8	29.3	5.8	2.1	4.3	6.1	1.7	3.4	3.6	6.5	5.0	3.5	3.6	3.5	6.8	9.4	19.6	6.1	9.1	6.4
BLOOMZ	3.8	4.6	2.8	5.3	2.9	4.1	5.1	3.4	4.2	32.3	4.9	3.0	1.5	2.4	1.5	4.0	2.7	5.7	1.2	3.2	4.9	2.6	4.6	3.3	24.1	5.4	10.1	5.7
박트리아노스-라마	10.9	11.6	8.9	12.3	8.8	22.0	32.1	8.5	12.1	25.1	11.4	6.9	6.4	8.2	4.1	10.9	8.7	14.1	4.3	8.4	15.2	8.0	11.4	10.8	19.4	16.6	23.4	12.6
AYA-23	9.3	10.5	8.0	11.6	6.9	14.2	17.5	5.6	8.3	18.3	11.3	5.7	4.0	5.9	2.7	8.1	7.6	12.2	3.3	9.0	8.8	6.5	10.4	6.8	24.3	10.6	17.7	9.8
AYA-101	26.4	26.8	14.6	21.6	12.6	49.3	46.6	33.3	25.8	49.5	38.8	12.2	25.9	37.2	4.4	17.8	13.4	29.7	17.6	13.2	23.3	20.4	35.6	22.2	36.5	36.9	41.9	27.2
바다사자	7.2	8.1	6.5	9.3	5.8	12.5	17.1	4.9	7.0	13.9	7.9	5.3	7.0	9.6	2.0	7.6	6.0	9.5	4.8	6.6	8.4	4.9	8.0	5.9	21.2	10.3	14.1	8.6
SeaLLM v2.5	15.2	20.2	11.7	19.5	11.5	37.1	49.1	14.5	26.8	43.0	26.6	7.5	17.8	22.2	4.7	15.1	12.2	26.8	9.2	14.6	19.2	9.4	22.0	21.6	36.7	28.8	45.7	21.8
선원	19.2	24.5	15.3	23.1	14.6	29.0	39.7	8.6	13.5	46.8	30.6	7.1	12.5	24.4	6.2	10.5	16.0	28.8	5.8	19.1	16.5	9.0	26.7	22.0	41.1	21.5	49.9	21.6
Cendol-mT5	8.3	11.4	14.2	11.6	6.9	7.2	8.4	4.7	5.5	35.8	17.5	4.0	6.3	8.5	2.0	5.2	6.1	10.5	2.9	8.8	6.6	4.1	17.1	5.5	4.4	6.4	20.5	9.3
Cendol-Llama2	8.6	10.0	14.4	19.3	6.6	6.9	8.2	6.4	6.4	36.1	19.1	5.5	3.0	4.3	4.1	4.5	14.1	22.0	1.9	17.5	5.4	4.8	17.3	3.4	8.1	7.6	22.0	10.6
메라크	7.4	10.3	6.7	11.3	7.1	8.2	12.8	6.3	6.7	29.5	9.6	3.7	3.8	5.9	3.2	8.0	6.5	12.5	2.4	8.0	8.2	5.6	10.6	5.9	7.2	7.4	20.4	8.7
왕찬X-라마3	19.8	24.4	14.3	28.9	13.4	42.2	48.6	12.7	29.4	50.1	29.4	7.7	18.1	19.7	6.0	17.6	15.6	30.0	10.4	18.1	22.4	13.9	28.0	25.1	39.2	35.5	45.4	24.7
말레이시아 라마3	15.2	17.3	12.3	22.2	11.1	19.7	24.0	8.7	12.6	38.6	19.4	7.2	6.7	9.0	5.9	10.6	12.4	23.5	4.2	14.3	13.9	8.3	19.0	14.2	17.3	15.6	44.4	15.8
전체	13.3	16.1	11.4	17.2	9.9	24.4	29.3	11.8	16.0	35.1	20.0	6.7	9.7	13.3	4.2	11.5	10.9	19.8	6.4	12.3	14.2	8.0	18.5	13.1	25.2	20.3	30.4	15.9

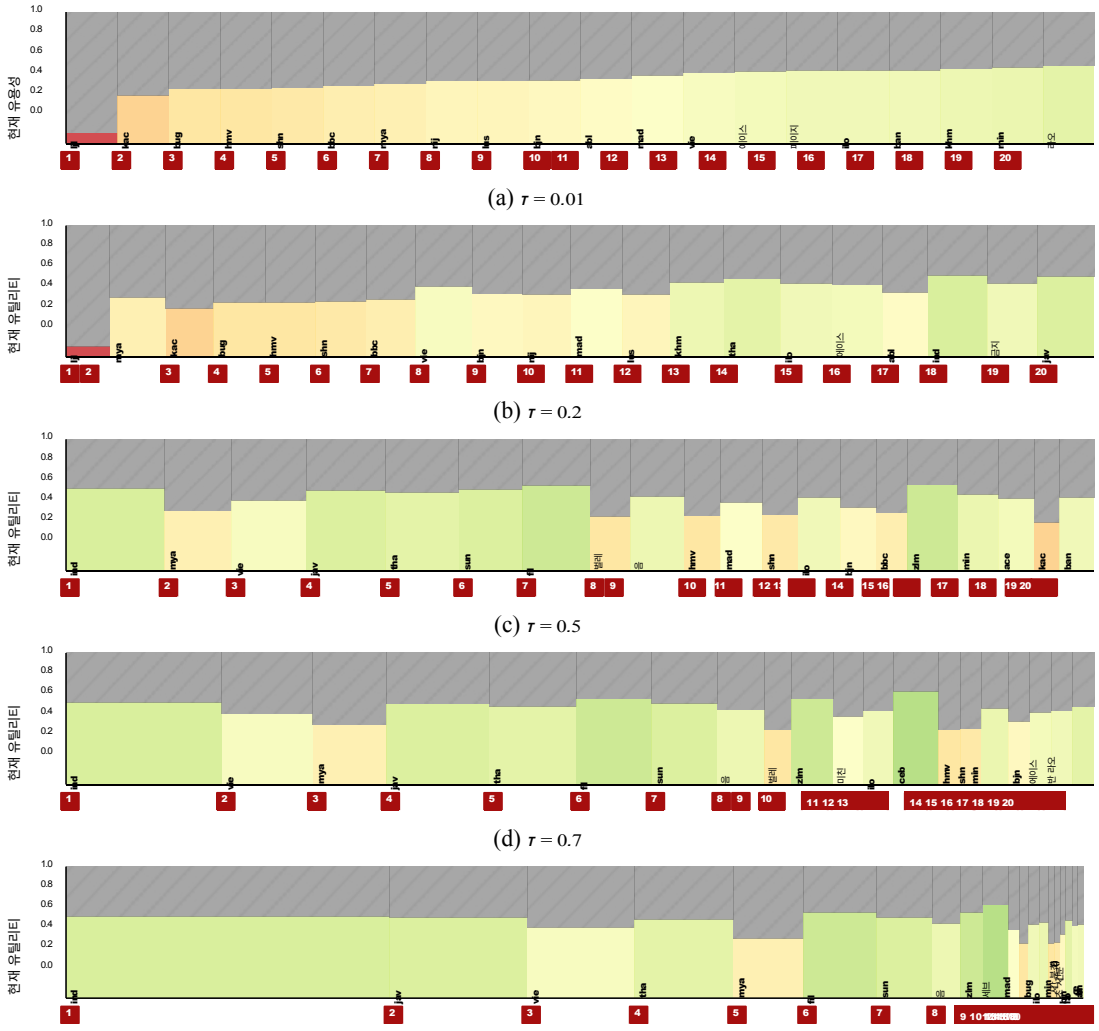
표 27: 언어별 ROUGE-L 기준 NLG 평가 결과.

언어	하위 집합	원본 작업	도메인	# 샘플
		번역체		
ENG	emotes_3k_eng_seacrowd_t2t	상식적 추론	윤리	2000
ENG	aya_평가_스위트_eng_seacrowd_t2t	지침 조정	일반	400
IND	belebele_ind_latn_seacrowd_qa	QA	일반	1969
IND	parallel_asian_treebank_ind_eng_seacrowd_t2t	기계 번역	뉴스	31
IND	aya_evaluation_suite_ind_seacrowd_t2t	지시 조정	일반	4
IND	bactrian_x_id_seacrowd_t2t	명령어 튜닝	혼합, 다중 도메인, 위키백과	1972
IND	seaeval_cross_logiqa_ind_seacrowd_qa	상식 추론, QA	해설, 일반, 다중 도메인, 문화 및 유산	16
IND	seaeval_cross_mmlu_ind_seacrowd_qa	상식적 추론, QA	해설, 일반, 다중 영역, 문화 및 유산	8
KHM	belebele_khm_khmr_seacrowd_qa	QA	일반	399
KHM	khmer_alt_pos_seacrowd_seq_label	POS 태깅	뉴스	1595
KHM	parallel_asian_treebank_khm_eng_seacrowd_t2t	기계 번역	뉴스	6
KHM	aya_evaluation_suite_khm_seacrowd_t2t	지시 조정	일반	8
KHM	bactrian_x_km_seacrowd_t2t	명령어 튜닝	혼합, 다중 도메인, 위키백과	1992
LAO	belebele_lao_laoo_seacrowd_qa	QA	일반	1969
라오스어	parallel_asian_treebank_lao_eng_seacrowd_t2t	기계 번역	뉴스	31
라오스어	aya_evaluation_suite_lao_seacrowd_t2t	지시 조정	일반	400
MYA	belebele_mya_mymr_seacrowd_qa	QA	일반	1969
MYA	parallel_asian_treebank_mya_eng_seacrowd_t2t	기계 번역	뉴스	31
MYA	aya_evaluation_suite_mya_seacrowd_t2t	지시 조정	일반	8
MYA	bactrian_x_my_seacrowd_t2t	명령어 튜닝	혼합, 다중 도메인, 위키백과	1992
FIL	belebele_tgl_latn_seacrowd_qa	QA	일반	2000
FIL	bactrian_x_tl_seacrowd_t2t	지시 조정	혼합, 다중 도메인, 위키백과	2000
THA	belebele_tha_thai_seacrowd_qa	QA	일반	1969
THA	parallel_asian_treebank_tha_eng_seacrowd_t2t	기계 번역	뉴스	31
THA	aya_evaluation_suite_tha_seacrowd_t2t	지시 조정	일반	8
THA	bactrian_x_th_seacrowd_t2t	명령어 튜닝	혼합, 다중 도메인, 위키백과	1992
VIE	belebele_vie_latn_seacrowd_qa	QA	일반	1969
VIE	parallel_asian_treebank_vie_eng_seacrowd_t2t	기계 번역	뉴스	31
VIE	aya_evaluation_suite_vie_seacrowd_t2t	지침 조정	일반	4
VIE	bactrian_x_vi_seacrowd_t2t	지시문 조정	혼합, 다중 도메인, 위키백과	1972
VIE	seaeval_cross_logiqa_vie_seacrowd_qa	상식 추론, QA	해설, 일반, 다중 도메인, 문화 및 유산	16
VIE	seaeval_cross_mmlu_vie_seacrowd_qa	상식적 추론, QA	해설, 일반, 다중 영역, 문화 및 유산	8
ZLM	belebele_zsm_latn_seacrowd_qa	QA	일반	1969
ZLM	parallel_asian_treebank_zlm_eng_seacrowd_t2t	기계 번역	뉴스	31
ZLM	aya_평가_스위트_zsm_seacrowd_t2t	지시 조정	일반	400
ZLM	seaeval_cross_logiqa_zlm_seacrowd_qa	상식적 추론, QA	해설, 일반, 다중 영역, 문화 및 유산	1056
ZLM	seaeval_cross_mmlu_zlm_seacrowd_qa	상식적 추론, QA	해설, 일반, 다중 영역, 문화 및 유산	300
		자연어		
ENG	cosem_seacrowd_ssp	언어 모델링	소셜 미디어	2000
IND	sea_bench_ind_seacrowd_t2t	지시 조정	논평, 일반, 다중 영역, 문화 및 유산	200
KHM	gklmip_newsclass_seacrowd_text	감정 분석	전자 상거래	1436
KHM	sea_bench_khm_seacrowd_t2t	지시 조정	해설, 일반, 다중 도메인, 문화 및 유산	160
LAO	sea_bench_lao_seacrowd_t2t	지침 조정	해설, 일반, 다중 영역, 문화 및 유산	160
MYA	gklmip_sentiment_seacrowd_text	감정 분석	전자 상거래	716
MYA	sea_bench_mya_seacrowd_t2t	지시 조정	해설, 일반, 다중 도메인, 문화 및 유산	160
FIL	sea_bench_tgl_seacrowd_t2t	지침 조정	해설, 일반, 다중 영역, 문화 및 유산	160
THA	sea_bench_tha_seacrowd_t2t	지침 조정	해설, 일반, 다중 영역, 문화 및 유산	40
THA	vistec_tp_th_21_seacrowd_seq_label	NER	소셜 미디어	1960
VIE	sea_bench_vie_seacrowd_t2t	지침 조정	논평, 일반, 다중 영역, 문화 및 유산	200
ZLM	sea_bench_zlm_seacrowd_t2t	지침 조정	해설, 일반, 다중 영역, 문화 및 유산	160

표 28: 번역어 분류기 실험에 사용된 훈련 데이터.

언어	하위 집합	원본 작업	도메인	샘플 수
		번역체		
ENG	emotes_3k_eng_seacrowd_t2t	상식적 추론	윤리	2000
ENG	aya_평가_스위트_eng_seacrowd_t2t	지침 조정	일반	400
IND	belebele_ind_latn_seacrowd_qa	QA	일반	1969
IND	parallel_asian_treebank_ind_eng_seacrowd_t2t	MT	뉴스	31
IND	aya_evaluation_suite_ind_seacrowd_t2t	지시 조정	일반	4
IND	bactrian_x_id_seacrowd_t2t	명령어 튜닝	혼합, 다중 도메인, 위키백과	1972
IND	seaeval_cross_logiqa_ind_seacrowd_qa	상식 추론, QA	해설, 일반, 다중 도메인, 문화 및 유산	16
IND	seaeval_cross_mmlu_ind_seacrowd_qa	상식적 추론, QA	해설, 일반, 다중 영역, 문화 및 유산	8
KHM	belebele_khm_khmr_seacrowd_qa	QA	일반	399
KHM	khmer_alt_pos_seacrowd_seq_label	POS 태깅	뉴스	1595
KHM	parallel_asian_treebank_khm_eng_seacrowd_t2t	MT	뉴스	6
KHM	aya_evaluation_suite_khm_seacrowd_t2t	지시 조정	일반	8
KHM	bactrian_x_km_seacrowd_t2t	명령어 튜닝	혼합, 다중 도메인, 위키백과	1992
LAO	belebele_lao_laoo_seacrowd_qa	QA	일반	1969
라오스어	parallel_asian_treebank_lao_eng_seacrowd_t2t	MT	뉴스	31
라오스어	aya_evaluation_suite_lao_seacrowd_t2t	지침 조정	일반	400
MYA	belebele_mya_mymr_seacrowd_qa	QA	일반	1969
MYA	parallel_asian_treebank_mya_eng_seacrowd_t2t	MT	뉴스	31
MYA	aya_evaluation_suite_mya_seacrowd_t2t	지시 조정	일반	8
MYA	bactrian_x_my_seacrowd_t2t	명령어 튜닝	혼합, 다중 도메인, 위키백과	1992
FIL	belebele_tgl_latn_seacrowd_qa	QA	일반	2000
FIL	bactrian_x_tl_seacrowd_t2t	지시 조정	혼합, 다중 도메인, 위키백과	2000
THA	belebele_tha_thai_seacrowd_qa	QA	일반	1969
THA	parallel_asian_treebank_tha_eng_seacrowd_t2t	MT	뉴스	31
THA	aya_evaluation_suite_tha_seacrowd_t2t	지시 조정	일반	8
THA	bactrian_x_th_seacrowd_t2t	명령어 튜닝	혼합, 다중 도메인, 위키백과	1992
VIE	belebele_vie_latn_seacrowd_qa	QA	일반	1969
VIE	parallel_asian_treebank_vie_eng_seacrowd_t2t	MT	뉴스	31
VIE	aya_evaluation_suite_vie_seacrowd_t2t	지침 조정	일반	4
VIE	bactrian_x_vi_seacrowd_t2t	지침 조정	혼합, 다중 도메인, 위키백과	1972
VIE	seaeval_cross_logiqa_vie_seacrowd_qa	상식 추론, QA	해설, 일반, 다중 도메인, 문화 및 유산	16
VIE	seaeval_cross_mmlu_vie_seacrowd_qa	상식적 추론, QA	해설, 일반, 다중 영역, 문화 및 유산	8
ZLM	belebele_zsm_latn_seacrowd_qa	QA	일반	1969
ZLM	parallel_asian_treebank_zlm_eng_seacrowd_t2t	MT	뉴스	31
ZLM	aya_평가_스위트_zsm_seacrowd_t2t	지시 조정	일반	400
ZLM	seaeval_cross_logiqa_zlm_seacrowd_qa	상식 추론, QA	논평, 일반, 다중 영역, 문화 및 유산	1056
ZLM	seaeval_cross_mmlu_zlm_seacrowd_qa	상식적 추론, QA	해설, 일반, 다중 영역, 문화 및 유산	300
		자연어		
ENG	cosem_seacrowd_ssp	언어 모델링	소셜 미디어	2000
IND	sea_bench_ind_seacrowd_t2t	지침 조정	해설, 일반, 다중 도메인, 문화 및 유산	200
KHM	gklmip_newsclass_seacrowd_text	감정 분석	전자 상거래	1436
KHM	sea_bench_khm_seacrowd_t2t	지시 조정	해설, 일반, 다중 도메인, 문화 및 유산	160
LAO	sea_bench_lao_seacrowd_t2t	지침 조정	해설, 일반, 다중 영역, 문화 및 유산	160
MYA	gklmip_sentiment_seacrowd_text	감정 분석	전자 상거래	716
MYA	sea_bench_mya_seacrowd_t2t	지시 조정	해설, 일반, 다중 도메인, 문화 및 유산	160
FIL	sea_bench_tgl_seacrowd_t2t	지시 조정	논평, 일반, 다중 영역, 문화 및 유산	160
THA	sea_bench_tha_seacrowd_t2t	지침 조정	해설, 일반, 다중 영역, 문화 및 유산	40
THA	vistec_tp_th_21_seacrowd_seq_label	NER	소셜 미디어	1960
VIE	sea_bench_vie_seacrowd_t2t	지침 조정	해설, 일반, 다중 영역, 문화 및 유산	200
ZLM	sea_bench_zlm_seacrowd_t2t	지침 조정	해설, 일반, 다중 영역, 문화 및 유산	160

표 29: 번역어 분류기 실험에 사용된 테스트 데이터.



(a) $\tau = 0.01$

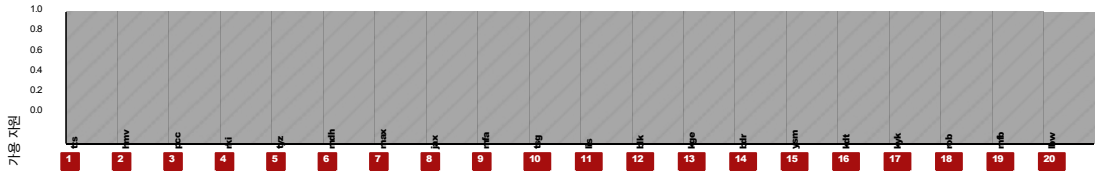
(b) $\tau = 0.2$

(c) $\tau = 0.5$

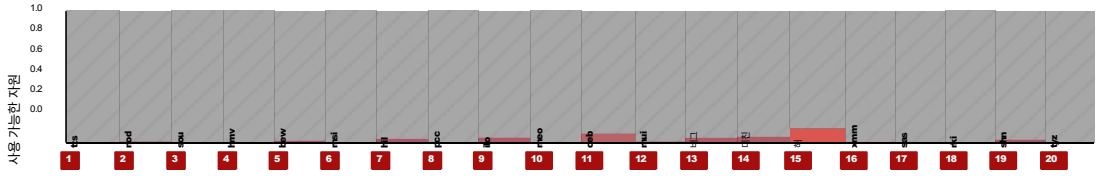
(d) $\tau = 0.7$

(e) $\tau = 1.0$

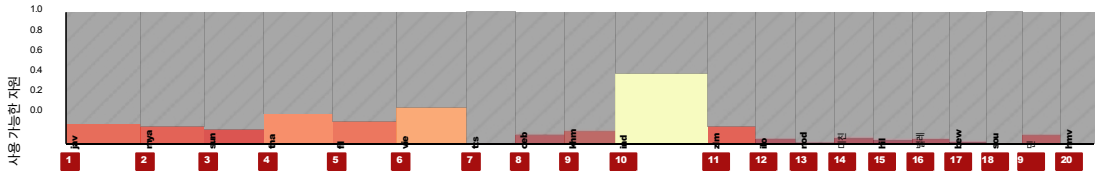
그림 10: 잠재적 수요와 현재 활용도를 기준으로 우선순위가 부여된 상위 20개 동남아시아 원주민 언어.



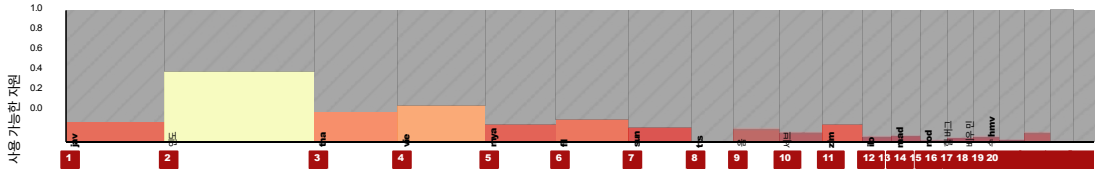
(a) $\tau = 0.01$



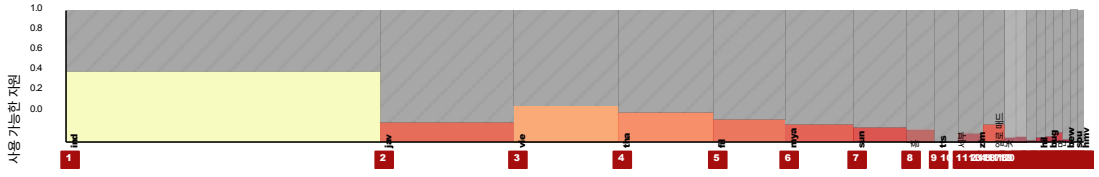
(b) $\tau = 0.2$



(c) $\tau = 0.5$



(d) $\tau = 0.7$



(e) $\tau = 1.0$

그림 11: 잠재적 수요와 데이터 가용성을 기준으로 우선 순위가 지정된 상위 20개 SEA 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인파		
1	인도네시아	인도네시아어	인도네시아	<1B
2	JAV	자바어	인도네시아	1억 미만
3	VIE	베트남어	베트남	<100M
4	THA	태국어	태국, 캄보디아	1억 미만
5	필리핀	필리핀	필리핀	1억 미만
6	MYA	버마어	미얀마	1억 미만
7	SUN	순다	인도네시아	<100M
8	TGL	타갈로그어	필리핀	<100M
9	KHM	크메르어	캄보디아, 베트남	1억 미만
10	CEB	세부아노어	필리핀	1억 미만
11	TTS	태국 북동부	태국	1억 미만
12	ZLM	말레이	말레이시아	1억 미만
13	ZSM	표준 말레이어	말레이시아, 브루나이, 싱가포르	1억 미만

표 30: 1천만 명 이상 사용자를 가진 동남아시아 원주민 언어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인구		
1	ILO	일로카노어	필리핀	1천만 미만
2	MAD	마두라	인도네시아	1천만 미만
3	NOD	태국 북부	라오스, 태국	1천만 미만
4	HIL	힐리가이논어	필리핀	1천만 미만
5	MIN	미낭카바우	인도네시아	1천만 미만
6	BUG	부기스	인도네시아	1천만 미만
7	BEW	베타위	인도네시아	<10M
8	SOU	태국 남부	태국	1천만 미만
9	라오스	라오스	캄보디아, 라오스	1천만 미만
10	HMV	몽어	베트남	1천만 미만
11	ACE	아체	인도네시아	1,000만 미만
12	BJN	반자르	인도네시아	1,000만 미만
13	BAN	발리	인도네시아	1,000만 미만
14	SHN	샨	미얀마, 태국	1,000만 미만
15	MUI	무시	인도네시아	1,000만 미만
16	MSI	사바 말레이	말레이시아	1천만 미만
17	MEO	케다 말레이	말레이시아, 태국	1,000만 미만
18	PCC	Giáy	베트남	<10M
19	WAR	와레이와레이어	필리핀	1천만 미만
20	MAK	마카사르	인도네시아	1,000만 미만
21	BCL	중부 비콜	필리핀	1,000만 미만
22	XMM	마나도 말레이	인도네시아	<10M
23	SAS	사삭어	인도네시아	1천만 미만
24	BBC	바탁 토바	인도네시아	1천만 미만
25	PAM	카팜팡간어	필리핀	1천만 미만
26	RKI	라카인	미얀마	1천만 미만
27	TYZ	타이족	베트남	1천만 미만
28	ABS	암본 말레이어	인도네시아	<10M
29	PSE	중앙 말레이	인도네시아	1천만 미만
30	IBA	이반	브루나이, 인도네시아, 말레이시아	1천만 미만
31	KXM	북부 크메르	태국	<10M
32	KHG	캄스 티베트어	미얀마	1천만 미만
33	KSW	S'gaw Karen	미얀마, 태국	1천만 미만
34	BTD	바탁 다이라이	인도네시아	1천만 미만
35	BTS	바탁 시말룽군	인도네시아	1천만 미만
36	CBK	차바카노어	필리핀	1천만 미만
37	PAG	팡가시난	필리핀	1천만 미만
38	MTQ	무옹	베트남	1천만 미만
39	BTM	바탁 만다일링	인도네시아	<10M
40	MDH	마긴다나온	필리핀	<10M
41	PMY	파푸아 말레이어	인도네시아	1천만 미만
42	GOR	고론탈로	인도네시아	1천만 미만
43	JAX	잠비 말레이	인도네시아	1천만 미만
44	KJP	동부 카렌족	미얀마, 태국	1천만 미만
45	MAX	북 올루카 말레이어	인도네시아	1천만 명 미만
46	MFA	파타니 말레이어	태국	<10M
		동남아시아에 속하지 않음Crowd		
47	MEP	마카사르 인도네시아어	인도네시아	1천만 미만

표 31: SEA 원주민 언어 중 사용자가 1천만 명 미만인 언어.

	ISO 639-3	언어	지역	인구
		동남아시아 크라우드		
1	NUT	농	베트남	<1M
2	KAC	징포	미얀마	100만 미만
3	TSG	타우수그어	필리핀	100만 미만
4	NJI	Ngaju	인도네시아	100만 미만
5	LJP	Lampung Api	인도네시아	<1M
6	MQY	망가라이	인도네시아	100만 미만
7	MRW	마라나오	필리핀	100만 미만
8	LJA	니아스	인도네시아	100만 미만
9	AKB	바탁 앙콜라	인도네시아	<1M
10	SDA	토라자-사단	인도네시아	100만 미만
11	MNW	몬	미얀마, 태국	100만 미만
12	HNI	하니	라오스, 베트남	100만 미만
13	KJG	크무족	라오스, 태국, 베트남	100만 미만
14	AOZ	우아브 메토	인도네시아	<1M
15	BLT	타이담	라오스, 베트남	100만 미만
16	LUS	미조 친	미얀마	100만 미만
17	CPS	카피즈어	필리핀	100만 미만
18	BTX	바탁 카로	인도네시아	100만 미만
19	LIS	리수	미얀마	100만 미만
20	MSB	마스바테뇨	필리핀	100만 미만
21	BLK	파오	미얀마, 태국	100만 미만
22	TDD	타이 누아	미얀마	100만 미만
23	DAY	랜드 다약	인도네시아	100만 미만
24	XDY	말레이계 다약족	인도네시아	100만 미만
25	BHP	비마	인도네시아	100만 미만
26	IBG	이바나그	필리핀	100만 미만
27	ZMI	네게리 셈빌란 말레이	말레이시아	100만 미만
28	MDR	만다린	인도네시아	100만 미만
29	KGE	코메링	인도네시아	100만 미만
30	BDR	서부 해안 바자우	말레이시아	100만 미만
31	KDT	쿠아이	캄보디아, 라오스, 태국	100만 미만
32	PRK	파라우크 와	미얀마	100만 미만
33	SGD	수리가오노	필리핀	100만 미만
34	TET	테툼어	동티모르, 인도네시아	<1M
35	BTO	린코나다 비콜	필리핀	100만 미만
36	TDT	테툼 달리	동티모르	<100만
37	IUM	유미엔	라오스, 베트남	100만 미만
38	KRJ	카나리야	필리핀	100만 미만
39	KYK	카마요	필리핀	100만 미만
40	LEW	레도 카일리	인도네시아	100만 미만
41	MKN	쿠팡 말레이	인도네시아	<1M
42	REJ	레장	인도네시아	<1M
43	MFB	방카	인도네시아	100만 미만
44	ROB	태국	인도네시아	<100만
45	LBW	Tolaki	인도네시아	100만 미만
46	KNX	켄다얀	인도네시아, 말레이시아	100만 미만
47	GAY	가요	인도네시아	100만 미만
48	MNB	무나	인도네시아	100만 미만
49	RBL	미라야 비콜	필리핀	100만 미만
50	SMW	Sumbawa	인도네시아	100만 미만
51	KXD	브루나이	브루나이	100만 미만
52	KHB	Lü	라오스, 미얀마	100만 미만
53	LHU	라후족	라오스, 미얀마	100만 미만
54	TWH	타이돈	라오스, 베트남	100만 미만
55	YSM	미얀마 수화	미얀마	100만 미만
56	DTP	카다잔 두순	말레이시아	100만 미만
57	PBL	서부 말바이 비콜	필리핀	100만 미만
58	KVR	케린치	인도네시아	<1M
59	PCE	루칭 팔라웅	미얀마	100만 미만
60	MRY	만다야	필리핀	100만 미만
61	NBE	코냐크 나가	미얀마	100만 미만
62	TCZ	타도 친	미얀마	100만 미만
63	JRA	자라이	캄보디아, 베트남	100만 미만
64	XBR	캄베라	인도네시아	100만 미만
65	MOG	몽곤도우	인도네시아	100만 미만
66	PWO	Pwo 서부 카렌	미얀마	100만 미만
67	CJA	서부 참	캄보디아, 베트남	100만 명 미만
68	AHK	아카	라오스, 미얀마, 태국	100만 미만
69	SSB	사마 남부	필리핀	100만 미만
70	SXN	상기르	인도네시아	100만 미만

표 32: (1/2) 사용자가 100만 명 미만인 동남아시아 토착어.

번	ISO 639-3	언어	지역	인구
		SEACrowd에서		
71	BTZ	바탁 알라스-클루엣	인도네시아	100만 미만
72	CTD	테딩 친	미얀마	100만 미만
73	SRV	남부 소르소가논	필리핀	100만 미만
74	ABL	람풍 노	인도네시아	100만 미만
75	DNW	서부 다니	인도네시아	100만 미만
76	KTP	카두오	라오스	100만 미만
77	SLP	라마홀롯	인도네시아	100만 미만
78	RAD	라데	베트남	<1M
79	SKI	Sika	인도네시아	<1M
80	KPM	코호	베트남	100만 미만
81	BDQ	바나르	베트남	<1M
82	BDL	인도네시아 바자우어	인도네시아	<1M
83	BPR	코로나달 블라안	필리핀	100만 미만
84	CCP	차크마	미얀마	100만 미만
85	KNE	칸카나이	필리핀	100만 미만
86	KYU	서부 카야	미얀마	<1M
87	MHY	마아난어	인도네시아	<1M
88	TNT	통템보안	인도네시아	100만 미만
89	PLL	슈에 팔라웅	미얀마	100만 미만
90	DAW	다와웨노	필리핀	100만 미만
91	CNH	하카 친	미얀마	100만 미만
92	SYB	중앙 수바넨	필리핀	100만 미만
93	RBB	루마이 팔라웅	미얀마	100만 미만
94	PMF	파모나	인도네시아	100만 미만
95	BLN	남부 카탄두아네스 비콜	필리핀	100만 미만
96	ITV	이타윗	필리핀	<1M
97	PDU	카얀	미얀마	100만 미만
98	MGM	만바에	동티모르	100만 미만
99	BHQ	Tukang Besi South	인도네시아	<1M
100	SLY	셀라야르	인도네시아	100만 미만
101	MVP	두리	인도네시아	100만 미만
102	BGZ	방가이	인도네시아	<1M
103	KJC	해안 콘조	인도네시아	100만 미만
104	SUC	서부 수바논	필리핀	100만 미만
105	CYO	쿠요논	필리핀	100만 미만
106	KHC	철공소 북	인도네시아	100만 미만
107	LHI	라후 시	미얀마	100만 미만
108	MEL	중앙 멜라나우	말레이시아	100만 미만
109	IBL	이발로이	필리핀	100만 미만
110	END	엔데	인도네시아	<1M
111	HVN	Hawu	인도네시아	100만 미만
112	KKV	Kangean	인도네시아	100만 미만
113	YKA	야칸	필리핀	100만 미만
114	LJL	리오	인도네시아	100만 미만
115	MKZ	마카사에	동티모르	100만 미만
116	BKD	비누키드	필리핀	<1M
117	BKR	바콩파이	인도네시아	100만 미만
118	EKG	에카리	인도네시아	100만 미만
119	HNJ	Hmong Njua	라오스, 태국, 베트남	100만 미만
120	KAK	칼랑구아	필리핀	100만 미만
121	KKH	쿤	미얀마	100만 미만
122	LBX	로완간	인도네시아	100만 미만
123	MHX	라오 보	미얀마	100만 미만
124	MQJ	마마사	인도네시아	100만 미만
125	PSP	필리핀 수화	필리핀	100만 미만
126	TGN	탄다가논	필리핀	100만 미만
		동남아시아에 속하지 않음		
127	RHG	로힝야족	미얀마	100만 미만
128	PHT	푸타이	라오스, 태국, 베트남	100만 미만
129	TVN	타보얀	미얀마	100만 미만
130	OSI	오싱	인도네시아	100만 미만
131	ILP	이란	필리핀	100만 미만
132	KZS	수구트 두순	말레이시아	100만 미만
133	VKT	텡가롱 쿠타이 말레이	인도네시아	100만 미만
134	PHU	푸안	라오스, 태국	100만 미만
135	CSH	아소 친	미얀마	100만 미만
136	MLC	카오란	베트남	100만 미만
137	KJK	고지 콘조	인도네시아	<1M
138	LJW	콜롬비아	인도네시아	100만 미만
139	SSS	소	라오스, 태국	100만 미만
140	DNV	다누	미얀마	100만 미만
141	SDQ	세만당	인도네시아	100만 미만
142	TJL	타이 랭	미얀마	100만 명 미만

표 33: (2/2) 사용자가 100만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
1	ADR	아도나라	인도네시아	10만 미만
2	SED	세당	베트남	<100K
3	BLF	부올	인도네시아	<100K
4	TBL	Tboli	필리핀	10만 미만
5	HRE	Hre	베트남	<100K
6	ROL	Romblomanon	필리핀	<100K
7	AKL	아클라논어	필리핀	10만 미만
8	TDN	톤다노	인도네시아	10만 미만
9	BPS	사랑가니 블란	필리핀	<100K
10	KQR	키마라강	말레이시아	10만 미만
11	SML	중앙 사마	필리핀	10만 미만
12	TXS	톤세아	인도네시아	10만 미만
13	STB	북부 수바넨	필리핀	10만 미만
14	BKS	북부 소르소가노	필리핀	10만 미만
15	KEI	케이	인도네시아	10만 미만
16	KLG	타가카올로	필리핀	10만 미만
17	TLD	탈라우드	인도네시아	10만 미만
18	ATB	자이와	미얀마	10만 미만
19	SSE	발랑기니 사마	필리핀	100,000 미만
20	TES	텡거르	인도네시아	10만 미만
21	TYR	타이 댕	라오스, 베트남	10만 미만
22	CIA	시아시아	인도네시아	10만 미만
23	GBI	갈레라	인도네시아	<100K
24	OTD	오트 다놈	인도네시아	<100K
25	CTS	북부 카탄두아네스 비콜 필리핀		10만 미만
26	LOE	샬루안	인도네시아	<100K
27	BNO	반토아논	필리핀	10만 미만
28	CMR	Mro-Khimi	미얀마	10만 미만
29	UBL	부히논 비콜	필리핀	10만 미만
30	CIM	둥부 참어	베트남	<100K
31	BKX	바이케노	동티모르	<100K
32	AAZ	아마라시	인도네시아	<100K
33	BHW	비아크	인도네시아	<100K
34	KQE	칼라간	필리핀	10만 미만
35	XNN	북부 칸카나이	필리핀	10만 미만
36	XSB	삼발	필리핀	100K 미만
37	CFM	팔람 친	미얀마	10만 미만
38	LBL	리본 비콜	필리핀	10만 미만
39	WLO	월리오	인도네시아	10만 미만
40	BTH	비야타 비다유	인도네시아, 말레이시아	<100K
41	KEM	케막	동티모르, 인도네시아	<100K
42	RAW	라왕	미얀마	10만 미만
43	TFT	테르나테	인도네시아	10만 미만
44	ZOM	조	미얀마	10만 미만
45	CNK	쿠미 친	미얀마	10만 미만
46	MQX	마무주	인도네시아	10만 미만
47	MSM	아구산 마노보	필리핀	10만 미만
48	NST	탕상 나가	미얀마	10만 미만
49	NXG	Ngad'a	인도네시아	<100K
50	OBO	오보 마노보	필리핀	<100K
51	PWW	Pwo Northern Karen	태국	10만 미만
52	SYA	상	인도네시아	10만 미만
53	TOM	톰볼루	인도네시아	10만 미만
54	XML	말레이시아 수화	말레이시아	100K 미만
55	MBS	사랑가니 마노보	필리핀	10만 미만
56	MWV	멘타와이	인도네시아	10만 미만
57	MSK	만사카	필리핀	10만 미만
58	SMK	볼리나오	필리핀	10만 미만
59	BFN	부낙	동티모르, 인도네시아	10만 미만
60	BGI	바고보-클라타	필리핀	<100K
61	DRG	룽구스	말레이시아	10만 미만
62	KZF	다아 카일리	인도네시아	<100K
63	WEW	웨제와	인도네시아	<100K
64	ROG	북부 로클라이	베트남	<100K
65	ILK	보칼롯	필리핀	<100K
66	KTV	동부 카투	베트남	10만 미만
67	DNT	미드 그랜드 밸리 다니	인도네시아	<100K
68	FRD	포데이터	인도네시아	<100K
69	MBT	마티살루그 마노보	필리핀	10만 미만
70	NXE	나게	인도네시아	10만 미만
71	PTT	엔레캉	인도네시아	10만 명 미만

표 34: (1/5) 10만 명 미만의 사용자를 가진 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
72	TIY	테두라이	필리핀	10만 미만
73	TJG	툰중	인도네시아	10만 미만
74	WMM	마이와	인도네시아	10만 미만
75	SDO	부카르-사동 비다유	인도네시아, 말레이시아	<100K
76	KYP	강	라오스	<100K
77	TVO	티도레	인도네시아	<100K
78	HOS	호치민 수화 베트남		<100K
79	MHS	부루	인도네시아	<100K
80	STI	볼로 스팅	캄보디아, 베트남	<100K
81	범를	Lauje	인도네시아	<100K
82	BGS	타가바와	필리핀	10만 미만
83	SJM	마푼	필리핀	100,000 미만
84	BLR	블랑	미얀마, 태국	10만 미만
85	RGS	남부 로글라이	베트남	10만 명 미만
86	SMR	시메루에	인도네시아	<100K
87	CZT	조통 친	미얀마	<100K
88	KVQ	게바 카렌	미얀마	10만 미만
89	MTD	무알랑	인도네시아	10만 미만
90	XXK	케오	인도네시아	<100K
91	TKD	투쿠데데	동티모르	<100K
92	KIX	키암니옹 나가	미얀마	<100K
93	BSB	브루나이 비사야어	브루나이, 말레이시아	<100K
94	DAO	다이 친	미얀마	10만 미만
95	DDG	파탈루쿠	동티모르	10만 미만
96	MQN	모로네네	인도네시아	<100K
97	GES	게세르-고롬	인도네시아	10만 미만
98	PHO	푸노이	라오스	<100K
99	SLM	팡구타란 사마	필리핀	10만 미만
100	HRO	하로이	베트남	10만 미만
101	IVV	이바탄	필리핀	10만 미만
102	MRH	마라 친	미얀마	10만 미만
103	BTW	부투아논	필리핀	10만 미만
104	CMA	마아	베트남	10만 미만
105	SBL	보톨란 삼발	필리핀	10만 미만
106	CMO	중앙 몽족	캄보디아, 베트남	10만 미만
107	BLZ	발란탁	인도네시아	<100K
108	TPU	타무판	캄보디아	<100K
109	BLJ	볼루간	인도네시아	<100K
110	CGC	카가얀넌	필리핀	10만 미만
111	CLU	칼루야눈	필리핀	10만 미만
112	CML	코넥코넥	인도네시아	10만 미만
113	GAD	가당	필리핀	<100K
114	HLT	마투 친	미얀마	10만 미만
115	IFK	투왈리 이푸가오	필리핀	10만 미만
116	IFU	마요야오 이푸가오	필리핀	10만 미만
117	KNB	루부아간 칼링가	필리핀	10만 미만
118	KSX	케당	인도네시아	10만 미만
119	LCF	루부	인도네시아	<100K
120	LSI	라시드	미얀마	<100K
121	MBA	히가오논	필리핀	10만 미만
122	MNG	동부 므농	베트남	10만 미만
123	MRO	므루	미얀마	10만 미만
124	MTA	코타바토 마노보	필리핀	10만 미만
125	SET	센타니	인도네시아	10만 미만
126	TMN	타만	인도네시아	<100K
127	TWU	테르마누	인도네시아	<100K
128	TXM	토미니	인도네시아	<100K
129	ULM	울루만다	인도네시아	<100K
130	와우	와보니	인도네시아	10만 미만
131	SNE	바우 비다유	인도네시아, 말레이시아	10만 미만
132	TDF	탈리엥	라오스	10만 미만
133	LBO	라벤	라오스	<100K
134	ACN	응오창	미얀마	<100K
135	TLB	토벨로	인도네시아	<100K
136	IFA	암가나드 이푸가오	필리핀	<100K
137	ITD	남티둥	인도네시아, 말레이시아	10만 미만
138	PHA	파형	베트남	<100K
139	ATD	아타 마노보	필리핀	10만 미만
140	BRU	동부 브루	라오스, 베트남	10만 미만
141	KZP	카이디팡	인도네시아	<100K
142	ABX	이나박논	필리핀	10만 명 미만

표 35: (2/5) 사용자가 10만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
143	AOL	알로르	인도네시아	10만 미만
144	JMD	암데나	인도네시아	<100K
145	LAA	남부 수바넌	필리핀	<100K
146	LMY	람보야	인도네시아	<100K
147	TXE	토톨리	인도네시아	<100K
148	OYB	Oy	라오스	<100K
149	MLF	말	라오스, 태국	<100K
150	LND	룬다예	브루나이, 인도네시아, 말레이시아	10만 미만
151	PRH	포로하논	필리핀	10만 미만
152	BRB	브라오	캄보디아, 라오스, 베트남	10만 미만
153	LBN	Rmeet	라오스	10만 미만
154	ILM	이란	말레이시아	10만 미만
155	PIU	밤밤	인도네시아	10만 미만
156	VKL	쿨리수수	인도네시아	10만 미만
157	BLW	발랑가오	필리핀	10만 미만
158	BSY	사바 비사야	말레이시아	10만 미만
159	KBR	크롱	캄보디아	10만 미만
160	DTB	라복-키나바탕간 카다잔 말레이시아		10만 미만
161	AYZ	마이 브라트	인도네시아	<100K
162	BAC	바두이	인도네시아	<100K
163	BRV	서부 브루	라오스, 태국	<100K
164	BWP	만도보 비와	인도네시아	<100K
165	DNA	어퍼 그랜드 밸리 다니	인도네시아	<100K
166	DNI	로어 그랜드 밸리 다니	인도네시아	<100K
167	DTR	로투드	말레이시아	10만 미만
168	DUN	두순 데야	인도네시아	<100K
169	KJE	카사르	인도네시아	<100K
170	KLI	칼롬팡	인도네시아	10만 미만
171	KOD	코디	인도네시아	<100K
172	LLG	Lole	인도네시아	<100K
173	LRT	라란투카 말레이	인도네시아	<100K
174	MNZ	모니	인도네시아	<100K
175	PEA	페라나칸 인도네시아인	인도네시아	<100K
176	PPK	우마	인도네시아	<100K
177	PRT	프라	라오스, 태국	10만 미만
178	TMM	타이탄	베트남	<100K
179	TNW	톤사왕	인도네시아	10만 미만
180	TWY	타워얀	인도네시아	10만 미만
181	TXQ	티이	인도네시아	10만 미만
182	WLW	Walak	인도네시아	10만 미만
183	SKH	시콜레	인도네시아	10만 미만
184	LHK	중앙 본톡	필리핀	10만 미만
185	CJE	Chru	베트남	10만 미만
186	HNN	하누누	필리핀	10만 미만
187	TLU	톨레후	인도네시아	<100K
188	WMH	와이마아	동티모르	10만 미만
189	HRK	하루쿠	인도네시아	100,000 미만
190	LEX	루앙	인도네시아	100,000 미만
191	푸오	푸옥	베트남	<100K
192	REN	Rengao	베트남	100,000 미만
193	ALP	알루네	인도네시아	10만 미만
194	BWE	Bwe Karen	미얀마	10만 미만
195	TLT	수 나마	인도네시아	100,000 미만
196	ZYP	자프 친	미얀마	10만 미만
197	ABZ	아부이	인도네시아	10만 미만
198	AKG	아나칼랑구	인도네시아	10만 미만
199	HAD	하팀	인도네시아	10만 미만
200	HTU	히투	인도네시아	<100K
201	NLC	날카	인도네시아	10만 미만
202	PAC	파코	라오스, 베트남	10만 미만
203	YOG	Yogad	필리핀	10만 미만
204	MXD	모당	인도네시아	<100K
205	JEH	Jeh	라오스, 베트남	<100K
206	KYN	북부 비누키드논	필리핀	<100K
207	PHG	Phuong	베트남	<100K
208	AGN	아구타옌	필리핀	10만 미만
209	CNW	Ngawn Chin	미얀마	<100K
210	ILA	일레 이페	인도네시아	<100K
211	KBD	카이루이-미디키	동티모르	10만 미만
212	LOA	로로다	인도네시아	<100K
213	MBB	서부 부키드논 마노보	필리핀	10만 미만
214	MWQ	뮈인 친	미얀마	10만 미만
215	NXA	나우에테	동티모르	10만 미만
216	PRF	파라난	필리핀	10만 명 미만

표 36: (3/5) 사용자가 10만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
217	SNL	상길	필리핀	<100K
218	TBY	타바루	인도네시아	10만 미만
219	TEA	테미아르	말레이시아	10만 미만
220	YLI	앙구루크 얄리	인도네시아	<100K
221	MEJ	메야	인도네시아	<100K
222	MBI	일리아넨 마노보	필리핀	10만 미만
223	PLW	브룩스 포인트 팔라와노	필리핀	10만 미만
224	DUU	드렁	미얀마	10만 미만
225	HEG	헬롱	인도네시아	10만 미만
226	MZQ	모리 아타스	인도네시아	10만 미만
227	UHN	다얄	인도네시아	10만 미만
228	XMZ	모리 바와	인도네시아	10만 미만
229	KJM	캄보디아	베트남	<100K
230	HAL	살랑	라오스, 베트남	10만 미만
231	IDT	이다테	동티모르	10만 미만
232	DOK	돈도	인도네시아	10만 미만
233	GAL	칼로렌	동티모르, 인도네시아	<100K
234	KSC	남부 칼링가	필리핀	10만 미만
235	TXA	통보누오	말레이시아	10만 미만
236	NGT	크리앙	라오스	10만 미만
237	KMK	리모스 칼링가	필리핀	100,000 미만
238	ALO	라리케-와카시후	인도네시아	10만 미만
239	YNO	용	태국	10만 미만
240	RIL	리앙 랑	미얀마	<100K
241	ATQ	아랄레-타불라한	인도네시아	10만 미만
242	CEK	동부 쿠미 친	미얀마	<100K
243	CUA	쿠아	베트남	100,000 미만
244	MNX	Sougb	인도네시아	100,000 미만
245	MQS	서 마키안	인도네시아	10만 미만
246	NUF	누수	미얀마	10만 미만
247	PLC	중앙 팔라완	필리핀	100,000 미만
248	PLV	남서부 팔라완	필리핀	10만 미만
249	RGU	리코우	인도네시아	10만 미만
250	SZW	사와이	인도네시아	10만 미만
251	TDJ	타지오	인도네시아	10만 미만
252	XKL	주류 케나어	인도네시아, 말레이시아	10만 미만
253	YIN	리앙 라이	미얀마	10만 미만
254	LCL	리셀라	인도네시아	100,000 미만
255	LRA	라라 바카티	인도네시아, 말레이시아	10만 미만
256	BVE	베라우 말레이	인도네시아	<100K
257	KML	타누단 칼링가	필리핀	10만 미만
258	BEU	Blagar	인도네시아	<100K
259	XEM	Mateq	인도네시아	<100K
260	LEV	서부 판타르	인도네시아	100,000 미만
261	PTN	파타니	인도네시아	10만 미만
262	OOG	Ong	라오스	10만 미만
263	SPR	사파루아	인도네시아	<100K
264	AMK	암바이	인도네시아	10만 미만
265	IFB	바타드 이푸가오	필리핀	10만 미만
266	AAX	만도보 아타스	인도네시아	10만 미만
267	BEP	베호아	인도네시아	10만 미만
268	BVY	바야바야논	필리핀	100,000명 미만
269	CSY	시인 친	미얀마	10만 미만
270	DBJ	이다안	말레이시아	10만 미만
271	EMB	엠발로	인도네시아	<100K
272	IRY	이라야	필리핀	10만 미만
273	JAK	자쿤	말레이시아	10만 미만
274	JAQ	야카이	인도네시아	10만 미만
275	KPS	테히트	인도네시아	<100K
276	KVB	쿠부	인도네시아	100,000 미만
277	KXF	카위아우	미얀마	10만 미만
278	KYT	카야가르	인도네시아	10만 미만
279	LJE	람피	인도네시아	<100K
280	LUR	Loura	인도네시아	10만 미만
281	MBD	디바바원 마노보	필리핀	10만 미만
282	MBF	바바 말레이	싱가포르	100K 미만
283	MKY	이스트 마키안	인도네시아	10만 미만
284	MVD	만보루	인도네시아	<100K
285	NDX	응두가	인도네시아	10만 미만
286	PEZ	동부 페난	브루나이, 말레이시아	10만 명 미만
287	PLE	팔루에	인도네시아	10만 미만
288	SEA	세마이	말레이시아	10만 미만
289	SSQ	소아	인도네시아	10만 명 미만

표 37: (4/5) 사용자가 10만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
290	SZB	Ngalum	인도네시아	<100K
291	TBK	칼라미안 타그반와	필리핀	10만 미만
292	TBW	타그반와	필리핀	10만 미만
293	TCX	타타나	말레이시아	10만 미만
294	WNK	와누카카	인도네시아	10만 미만
295	YVA	야와	인도네시아	10만 미만
		동남아시아 제외Cro	*wd*	
296	INT	인타	미얀마	10만 미만
297	LOC	이노한	필리핀	10만 미만
298	MQG	코타 방군 쿠타이 말레이	인도네시아	10만 미만
299	BFX	반타야논	필리핀	10만 미만
300	TOU	Tho	베트남	100K 미만
301	NCQ	북부 카탕	라오스	100,000 미만
302	BVU	부킷 말레이	인도네시아	10만 미만
303	BYD	베냐두	인도네시아	10만 미만
304	TSQ	태국 수화	태국	<100K
305	NYW	Nyaw	태국	10만 미만
306	RIR	리본	인도네시아	<100K
307	SCG	상가우	인도네시아	10만 미만
308	SCT	남부 카탕	라오스	<100K
309	STT	부데 스텅	베트남	10만 미만
310	TCO	타웅요	미얀마	100,000 미만
311	VKK	카우르	인도네시아	10만 미만
312	HAB	하노이 수화	베트남	<100K
313	DJO	장강	인도네시아	10만 미만
314	SBX	세베루앙	인도네시아	10만 미만
315	LSO	라오스 수화	라오스	<100K
316	SEZ	센탕 친	미얀마	10만 미만
317	SQA	태국 노래	태국	10만 미만
318	KNL	케닌잘	인도네시아	10만 미만
319	TTH	어퍼 타오이	라오스, 베트남	10만 미만
320	APG	암파낭	인도네시아	<100K
321	MNN	남부 므농	베트남	10만 명 미만
322	PEL	페칼	인도네시아	100,000 미만
323	ZKD	카두	미얀마	100,000 미만
324	BKZ	붕쿠	인도네시아	<100K
325	MKX	키나미긴 마노보	필리핀	100,000 미만
326	BNU	벤통	인도네시아	10만 미만
327	KXY	카용	베트남	100K 미만
328	MHP	발리 말레이	인도네시아	10만 미만
329	UNZ	운데 카일리	인도네시아	10만 미만
330	BLD	볼랑고	인도네시아	10만 미만
331	KUF	서부 카투	라오스	<100K
332	DNK	Dengka	인도네시아	10만 미만
333	MVV	타갈 무루트	인도네시아, 말레이시아	10만 미만
334	SKN	콜리부간 수바논	필리핀	10만 미만
335	SZN	술라	인도네시아	10만 미만
336	CNB	우푸 친	미얀마	10만 미만
337	BHV	바하우	인도네시아	10만 미만
338	ITT	Maeng Itneg	필리핀	10만 미만
339	HJI	하지	인도네시아	10만 미만
340	GHK	게코 카렌	미얀마	10만 미만
341	KVL	카야우	미얀마	10만 미만
342	TTO	로어 타오이	라오스	10만 미만
343	BDB	바삿	인도네시아	10만 미만
344	CLJ	라이투 친	미얀마	10만 미만
345	CLT	라투 친	미얀마	10만 미만
346	DUP	두아노	인도네시아, 말레이시아	10만 미만
347	KYB	부트부트 칼링가	필리핀	10만 미만
348	STG	Trieng	베트남	100,000 미만
349	CBW	키나발리안	필리핀	10만 미만
350	CSV	섭투 친	미얀마	<100K
351	RIU	리옹	인도네시아	<100K
352	SRG	술로드	필리핀	10만 미만
353	ITY	모야단 이트네그	필리핀	10만 미만
354	KKG	마바카 밸리 칼링가	필리핀	10만 미만
355	BNE	빈타우나	인도네시아	10만 미만
356	NLK	니니아 얄리	인도네시아	10만 미만
357	HIK	세이트 카이테투	인도네시아	10만 미만
358	KSN	카시구라닌	필리핀	10만 미만
359	TSL	쯔롱라오	베트남	<100K
360	XAO	카오	베트남	10만 명 미만

표 38: (5/5) 사용자가 10만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
1	XTE	케탱반	인도네시아	10,000 미만
2	BNA	보네라테	인도네시아	<10K
3	BKU	부히드	필리핀	10,000 미만
4	AWS	사우스 아위유	인도네시아	10,000 미만
5	WOO	마놈바이	인도네시아	10,000 미만
6	ASC	카수아리나 해안 아스마트	인도네시아	10,000 미만
7	TIH	티무곤 무루트	말레이시아	10,000 미만
8	ASL	아실루루	인도네시아	10,000 미만
9	SGB	마그안치 아이타	필리핀	10,000 미만
10	EKY	동부 카야	미얀마, 태국	<10K
11	IFY	Keley-i Kallahan	필리핀	10,000 미만
12	INL	인도네시아 수화	인도네시아	<10K
13	KGQ	카모로어	인도네시아	10,000 미만
14	KHT	캄티	미얀마	10,000 미만
15	KPQ	코루펀-셀라	인도네시아	10,000 미만
16	KTI	북무유	인도네시아	10,000 미만
17	LCP	서부 라와	태국	10,000명 미만
18	MTJ	모스코나	인도네시아	10,000 미만
19	SLU	셀라루	인도네시아	10,000 미만
20	TMW	테우만	말레이시아	10,000 미만
21	TXT	Citak	인도네시아	<10K
22	WHK	와하우 케냐	인도네시아	10,000 미만
23	TXN	웨스트 타랑간	인도네시아	10,000 미만
24	DRO	다로-마투 멜라나우	말레이시아	<10K
25	AWU	중앙 아위우	인도네시아	<10K
26	ITB	비농간 이트네그	필리핀	10,000 미만
27	LTI	레티	인도네시아	10,000 미만
28	SAJ	사후	인도네시아	10,000 미만
29	KVV	콜라	인도네시아	10,000 미만
30	KVU	Yinbaw	미얀마	10,000 미만
31	AKC	Mpur	인도네시아	10,000 미만
32	CNS	중앙 아스마트	인도네시아	10,000 미만
33	CRW	Chrau	베트남	10,000 미만
34	LWL	동부 라와	태국	10,000명 미만
35	LZN	라농 나가	미얀마	10,000 미만
36	MRZ	마린드	인도네시아	10,000 미만
37	ROW	Dela-Oenale	인도네시아	10,000 미만
38	SFE	동부 수바넨	필리핀	10,000 미만
39	TTD	투퉁	브루나이	10,000 미만
40	IWO	모로프	인도네시아	10,000 미만
41	TWB	타우부이드	필리핀	10,000 미만
42	BHZ	바다	인도네시아	<10K
43	PWM	몰보그	말레이시아, 필리핀	10,000 미만
44	PSA	아수 에와유	인도네시아	<10K
45	EBK	동부 본톡	필리핀	10,000 미만
46	TRE	동부 타랑간	인도네시아	10,000 미만
47	NPY	나푸	인도네시아	<10K
48	GDG	가당	필리핀	10,000 미만
49	GIR	레드 겔라오	베트남	<10K
50	KLL	카간 칼라간	필리핀	10,000 미만
51	LWT	레워토비	인도네시아	<10K
52	MOO	모노	베트남	10,000 미만
53	PNP	판카나	인도네시아	10,000 미만
54	TDR	토드라	베트남	<10K
55	WEO	와엘레	인도네시아	10,000 미만
56	WOI	카망	인도네시아	10,000 미만
57	WRP	와로펜	인도네시아	10,000 미만
58	LHA	라하	베트남	10,000 미만
59	KVO	도벨	인도네시아	10,000 미만
60	MTG	우나	인도네시아	10,000 미만
61	INN	이시나이	필리핀	10,000 미만
62	IHP	이하	인도네시아	<10K
63	JKA	카에라	인도네시아	10,000 미만
64	MYL	모마	인도네시아	10,000 미만
65	MMN	미나만와	필리핀	10,000 미만
66	NXR	닝게룸	인도네시아	10,000 미만
67	BLX	마그-인디 아이타	필리핀	10,000 미만
68	DUW	두순 위투	인도네시아	<10K
69	KGW	카론 도리	인도네시아	10,000 미만
70	KYO	클론	인도네시아	<10K
71	LBT	라치	베트남	10,000 미만
72	MLI	말림푼	인도네시아	10,000 미만
73	NFA	다오	인도네시아	10,000 미만
74	PDO	파도	인도네시아	<10K
75	RAZ	라힘부	인도네시아	<10K
76	TPG	콜라	인도네시아	10,000 미만
77	URK	우락 라워이	태국	<10K
78	WAD	와메사	인도네시아	10,000 미만
79	WOD	Wolani	인도네시아	10,000 미만
80	WUL	실리모	인도네시아	10,000명 미만

표 39: (1/6) 사용자가 10,000명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
81	YAC	패스 밸리 알리	인도네시아	<10K
82	YOY	Yoy	라오스, 태국	1만 미만
83	AND	안수스	인도네시아	10,000 미만
84	MXN	모이 켈림	인도네시아	10,000 미만
85	TLV	탈리아부	인도네시아	10,000 미만
86	BTY	보봇	인도네시아	<10K
87	DUQ	두순 말랑	인도네시아	10,000 미만
88	UMS	펜다우	인도네시아	10,000 미만
89	VBB	동남 바바르	인도네시아	<10K
90	BAJ	바라카이	인도네시아	<10K
91	BGR	바움 친	미얀마	10,000 미만
92	IRR	이르	라오스	<10K
93	NBQ	응겜	인도네시아	10,000 미만
94	BQR	부루수	인도네시아	<10K
95	KVD	쿠이	인도네시아	10,000 미만
96	BNY	빈툴루	말레이시아	10,000 미만
97	RKA	크라올	캄보디아	10,000 미만
98	JAH	자 헛	말레이시아	10,000 미만
99	KYS	바람 카얀	말레이시아	10,000 미만
100	SMU	솜레이	캄보디아	10,000 미만
101	SZA	세멜라이	말레이시아	10,000 미만
102	ALK	알락	라오스	<10K
103	ANL	아누콩소 친	미얀마	10,000 미만
104	BEI	바카티	인도네시아	<10K
105	IRH	이라루투	인도네시아	<10K
106	KTA	카투아	베트남	<10K
107	KTS	남무유	인도네시아	<10K
108	KZI	켈라빗	인도네시아, 말레이시아	10K
109	LMR	라마레라	인도네시아	<10K
110	MWT	모켄	미얀마, 태국	10,000명 미만
111	NTX	탕물 나가	미얀마	10,000명 미만
112	ROR	롱가	인도네시아	<10K
113	SDU	사루두	인도네시아	<10K
114	SLZ	마야	인도네시아	10,000 미만
115	SRE	사라 바카티	인도네시아	10,000 미만
116	TGB	토빌롱	말레이시아	10,000 미만
117	TWE	Teiwa	인도네시아	10,000 미만
118	TYN	콤바이	인도네시아	10,000 미만
119	WAH	와투벨라	인도네시아	<10K
120	NEV	냐헝	라오스	<10K
121	KLZ	카볼라	인도네시아	10,000 미만
122	AWY	에데라 아위	인도네시아	<10K
123	MQR	마니데	필리핀	10,000 미만
124	TNM	타볼라	인도네시아	10,000 미만
125	SKB	세크	라오스, 태국	10,000 미만
126	KVW	베르징	인도네시아	10,000 미만
127	XOD	코코다	인도네시아	<10K
128	BPQ	반다 말레이	인도네시아	<10K
129	BAY	바둘레이	인도네시아	<10K
130	KGX	카마루	인도네시아	<10K
131	KHE	코로와이	인도네시아	10,000 미만
132	LKJ	레문	말레이시아	10,000 미만
133	PKU	Paku	인도네시아	<10K
134	SAW	사위	인도네시아	<10K
135	TCG	타마가리오	인도네시아	10,000 미만
136	PNE	서부 페난	말레이시아	10,000 미만
137	XKS	콤베와하	인도네시아	10,000 미만
138	PGU	파구	인도네시아	<10K
139	TPO	타이 파오	라오스, 베트남	10,000 미만
140	ZRS	마이라시	인도네시아	10,000 미만
141	KZZ	칼라브라	인도네시아	<10K
142	BLS	발라에상	인도네시아	<10K
143	KUV	쿠르	인도네시아	<10K
144	REE	레장 카얀	말레이시아	10,000 미만
145	ABP	아벨렌 아이타	필리핀	10,000 미만
146	ADN	아당	인도네시아	10,000 미만
147	AHH	아구	인도네시아	<10K
148	BND	반다	인도네시아	10,000 미만
149	BNQ	반틱	인도네시아	10,000 미만
150	CKH	차크	미얀마	<10K
151	DUE	우마레이 두마겟 아그타	필리핀	1만 미만
152	EIP	좋아요	인도네시아	<10K
153	KGR	아본	인도네시아	10,000 미만
154	KIG	키마기마	인도네시아	<10K
155	NSY	나살	인도네시아	<10K
156	SWT	사윌라	인도네시아	10,000 미만
157	TMG	테르나테ɲc	인도네시아	10,000 미만
158	WMS	웜본	인도네시아	10,000 미만
159	MHE	마 메리	말레이시아	10,000 미만
160	BGL	보	라오스	10,000명 미만

표 40: (2/6) 사용자가 1만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
161	BPV	비안 마린드	인도네시아	10,000 미만
162	GZN	가네	인도네시아	<10K
163	DMR	동 다마르	인도네시아	10,000 미만
164	OBK	남부 본톡	필리핀	10,000 미만
165	BZL	보아노	인도네시아	<10K
166	HBU	하분	동티모르	<10K
167	ZNG	망	베트남	<10K
168	GEI	게베	인도네시아	<10K
169	SPB	Sepa	인도네시아	10,000 미만
170	AGV	레몬타도 두마갓	필리핀	10,000 미만
171	BZQ	불리	인도네시아	10,000 미만
172	BRP	Barapasi	인도네시아	10,000 미만
173	CBL	부알카우 친	미얀마	10,000 미만
174	GRS	그레시	인도네시아	10,000 미만
175	JMN	마쿠리 나가	미얀마	10,000 미만
176	KMT	켐투익	인도네시아	<10K
177	KWE	크웨르바	인도네시아	<10K
178	SKO	세코 텡가	인도네시아	<10K
179	WRS	와리스	인도네시아	10,000 미만
180	KYI	키푸트	말레이시아	10,000 미만
181	NRM	나롬	말레이시아	10,000 미만
182	KLW	타도	인도네시아	10,000 미만
183	SPU	사푸안	라오스	10,000 미만
184	JEI	예이	인도네시아	<10K
185	SQQ	수	라오스	10,000 미만
186	AWV	Jair Awyu	인도네시아	10,000 미만
187	BUP	부소아	인도네시아	10,000 미만
188	KKL	코사렉 예일	인도네시아	10,000 미만
189	ZKA	카임불라와	인도네시아	10,000 미만
190	KJR	쿠루두	인도네시아	<10K
191	ALJ	알랑간	필리핀	10,000 미만
192	ASY	야오사코르 아스마트	인도네시아	<10K
193	DMS	담펠라스	인도네시아	10,000 미만
194	ENR	에멩	인도네시아	10,000 미만
195	HNU	Hung	라오스, 베트남	10,000 미만
196	KWT	K웨스턴	인도네시아	10,000 미만
197	KYJ	Karao	필리핀	10,000 미만
198	LAU	Laba	인도네시아	10,000 미만
199	LEY	리몰라	인도네시아	10,000 미만
200	MQF	모무나	인도네시아	10,000 미만
201	MQO	모돌레	인도네시아	10,000 미만
202	NIR	님보란	인도네시아	<10K
203	PMO	품	인도네시아	10,000 미만
204	SGE	Segai	인도네시아	<10K
205	SZC	세막 베리	말레이시아	10,000 미만
206	TGT	중앙 타그반와	필리핀	10,000 미만
207	TTY	시카리타이	인도네시아	<10K
208	BGK	비트	라오스	<10K
209	GRM	코타 마루두 탈란탕	말레이시아	10,000 미만
210	SRL	이시라와	인도네시아	<10K
211	WBW	Woi	인도네시아	<10K
212	SIB	세뿡	말레이시아	10,000 미만
213	BNB	부칸 무루트	말레이시아	10,000 미만
214	LLM	Lasalimu	인도네시아	<10K
215	RMM	로마	인도네시아	10,000 미만
216	PCB	배	캄보디아	10,000 미만
217	ABC	앰발라 아이타	필리핀	10,000 미만
218	NXX	나프리	인도네시아	10,000 미만
219	LWH	화이트 라치	베트남	<10K
220	URY	오리아	인도네시아	<10K
221	IRX	캄베라우	인도네시아	10,000 미만
222	ATK	아티	필리핀	10,000 미만
223	BGB	보봉코	인도네시아	<10K
224	BVZ	바우지	인도네시아	<10K
225	BZP	켐베라노	인도네시아	<10K
226	CBN	나쿠르	태국	10,000 미만
227	DBF	에도피	인도네시아	<10K
228	ENO	엥가노	인도네시아	<10K
229	MKM	모클렌	태국	10,000 미만
230	NXL	사우스 누아울루	인도네시아	10,000 미만
231	VKO	코데오하	인도네시아	10,000 미만
232	WBB	와보	인도네시아	10,000 미만
233	YIR	North Awyu	인도네시아	<10K
234	ZBC	중앙 베라완	말레이시아	10,000 미만
235	BYA	바탁	필리핀	1만 명 미만

표 41: (3/6) 사용자가 1만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
236	BDG	봉기	말레이시아	<10K
237	FAU	파유	인도네시아	10,000 미만
238	ILU	일리운	인도네시아	<10K
239	YET	Yetfa	인도네시아	<10K
240	DMY	소와리	인도네시아	<10K
241	DDW	다웨라-다멜루르	인도네시아	10,000 미만
242	JHI	제하이	말레이시아	10,000 미만
243	XMT	Matbat	인도네시아	<10K
244	BEG	벨라이트	브루나이	10,000 미만
245	IVB	이바탄	필리핀	10,000 미만
246	OIA	오이라타	인도네시아	<10K
247	BKL	베릭	인도네시아	<10K
248	DUO	두파니난 아그타	필리핀	10,000 미만
249	KDW	코네라우	인도네시아	<10K
250	MSF	Mekwei	인도네시아	10,000 미만
251	NQM	Ndom	인도네시아	10,000 미만
252	SBG	모이 레마스	인도네시아	<10K
253	SEU	세루이-라우트	인도네시아	<10K
254	TVE	태운	인도네시아	<10K
255	TZN	Tugun	인도네시아	<10K
256	WNG	왕곰	인도네시아	<10K
257	BNJ	방곤	필리핀	10,000 미만
258	SNV	사반	인도네시아, 말레이시아	<10K
259	BDW	바하마	인도네시아	<10K
260	RAN	Riantana	인도네시아	10,000 미만
261	RNN	Roon	인도네시아	<10K
262	SZP	Suabo	인도네시아	<10K
263	ZBE	동 베라완	말레이시아	10,000 미만
264	SCB	쭈트	라오스, 베트남	10,000 미만
265	TVM	Tela-Masbuar	인도네시아	10,000 미만
266	UDJ	우지르	인도네시아	<10K
267	AGY	남부 알타	필리핀	10,000 미만
268	AIR	에어로란	인도네시아	10,000 미만
269	AQM	아토와임	인도네시아	10,000 미만
270	ASI	부루와임	인도네시아	<10K
271	ATT	팜플로나 아타	필리핀	10,000 미만
272	BCD	노스 바바르	인도네시아	10,000 미만
273	BNF	마시왕	인도네시아	<10K
274	BTQ	바텍	말레이시아	10,000 미만
275	CTH	타이푼 친	미얀마	10,000 미만
276	DEM	뎀	인도네시아	<10K
277	DMG	어퍼 키나바탕간	말레이시아	10,000 미만
278	DNU	다나우	미얀마	10,000 미만
279	ETZ	세미미	인도네시아	10,000 미만
280	JBJ	아란다이	인도네시아	10,000 미만
281	KBV	Dla	인도네시아	<10K
282	KPU	카포아	인도네시아	<10K
283	KVY	Yintale	미얀마	10,000 미만
284	MSG	모라이드	인도네시아	<10K
285	NKS	노스 아스마트	인도네시아	<10K
286	PNX	퐁크난	라오스	10,000 미만
287	SOB	소베이	인도네시아	<10K
288	WGO	암벨	인도네시아	<10K
289	WNO	와노	인도네시아	10,000 미만
290	XSE	Sempan	인도네시아	<10K
291	ZBW	웨스트 베라완	말레이시아	10,000 미만
		동남아시아에 속하지 않음Crowd		
292	RBK	북부 본톡	필리핀	10,000 미만
293	KVT	라타	미얀마	10,000 미만
294	LBG	라오팡	라오스	<10K
295	STU	삼타오	미얀마	10,000 미만
296	KXK	자이엔	미얀마	10,000 미만
297	ITI	인라우드 이트네그	필리핀	10,000 미만
298	NQQ	첸 카유 나가	미얀마	10,000 미만
299	PNC	파네이	인도네시아	10,000 미만
300	ZKN	카난	미얀마	10,000 미만
301	MLZ	말레이어	필리핀	<10K
302	KHF	쿠엔	라오스	<10K
303	KKX	코힌	인도네시아	<10K
304	LMJ	웨스트 렘바타	인도네시아	<10K
305	DKR	쿠이아우	말레이시아	10,000 미만
306	EBC	베긴치	인도네시아	<10K
307	MTW	남부 비누키드논	필리핀	10,000 미만
308	MQK	라자 카분수와 만오보	필리핀	10,000 미만
309	CSX	캄보디아 수화	캄보디아	10,000 미만
310	TIS	마사디잇 이트네그	필리핀	10,000 미만
311	CSJ	송라이 친	미얀마	10,000 미만
312	MQC	망골레	인도네시아	<10K
313	BPZ	빌바	인도네시아	<10K
314	LMF	남 렘바타	인도네시아	<10K
315	WHA	수 우파	인도네시아	10,000 미만
316	LKC	쿠쿵	베트남	10,000 미만
317	MQA	마바	인도네시아	10,000 미만
318	LCQ	루후	인도네시아	10,000 미만
319	MJB	마칼레로	동티모르	10,000명 미만

표 42: (4/6) 사용자가 1만 명 미만인 동남아시아 토착어.

번	ISO 639-3	언어	지역	인구
		SEACrowd에 포함되지 않음		
320	KRV	카뱃	캄보디아	<10K
321	CEY	에카이 친	미얀마	<10K
322	KJT	프레 포 카렌	태국	10,000 미만
323	KUK	케포	인도네시아	<10K
324	PUT	푸토	인도네시아	<10K
325	RIG	라종	인도네시아	<10K
326	SIB	사자우 바샷	인도네시아	10,000 미만
327	TKZ	타쿠아	베트남	10,000 미만
328	AMV	암벨라우	인도네시아	<10K
329	WLH	웰라운	동티모르, 인도네시아	10,000 미만
330	PLZ	팔루안 무루트	말레이시아	10,000 미만
331	JKP	파쿠 카렌	미얀마	10,000 미만
332	ADB	아타우란	동티모르	<10K
333	NEA	동부 응가다어	인도네시아	<10K
334	NTD	북부 티둥	말레이시아	10,000 미만
335	PHH	풀라	베트남	<10K
336	REB	렘봉	인도네시아	<10K
337	SKX	세코 파당	인도네시아	10,000 미만
338	SWU	수와와	인도네시아	<10K
339	TGR	타렝	라오스	10,000 미만
340	WEU	라웅투 친	미얀마	10,000 미만
341	사우디아라비아	살레만	인도네시아	10,000 미만
342	THI	타이 룽	라오스	10,000 미만
343	낮음	탐피아스 로부	말레이시아	10,000 미만
344	NPG	포뇨공왕 나가	미얀마	10,000 미만
345	UKK	무악 사악	미얀마	10,000 미만
346	TLQ	타이 로이	라오스, 미얀마	10,000 미만
347	HKN	멜-카오	캄보디아	10,000 미만
348	JKM	모브와 카렌	미얀마	10,000 미만
349	LMQ	라마투카	인도네시아	<10K
350	LVU	레부카	인도네시아	<10K
351	LWE	레우렝	인도네시아	<10K
352	RTC	룽투 친	미얀마	10,000 미만
353	RUU	라나스 로부	말레이시아	10,000 미만
354	TIU	아다센	필리핀	10,000 미만
355	UMN	파웅뉘안 나가	미얀마	<10K
356	LHH	라하	인도네시아	10,000 미만
357	BJX	바나우 칼링가	필리핀	10,000 미만
358	BVT	바티	인도네시아	<10K
359	KQV	오콜로드	인도네시아, 말레이시아	10,000 미만
360	XKK	카츰	캄보디아	10,000 미만
361	IWK	I-wak	필리핀	10,000 미만
362	LKA	라칼레이	동티모르	<10K
363	BZN	보아노	인도네시아	<10K
364	SBR	셈박옹 무루트	인도네시아, 말레이시아	10,000 미만
365	BFG	부상 카얀	인도네시아	<10K
366	HAP	후플라	인도네시아	<10K
367	KXI	케닝가우 무루트	말레이시아	10,000 미만
368	LLQ	로락	인도네시아	<10K
369	ROC	카크지아 로클라이	베트남	<10K
370	SLS	싱가포르 수화	싱가포르	<10K
371	STE	Liana-Seti	인도네시아	<10K
372	울루	우마 룽	인도네시아	<10K
373	WLI	와이올리	인도네시아	10,000 미만
374	WRX	와에 라나	인도네시아	<10K
375	XHV	쿠아	라오스, 베트남	10,000 미만
376	TDY	타디야완	필리핀	10,000 미만
377	ZBT	바투이	인도네시아	<10K
378	SWS	셀루와산	인도네시아	<10K
379	PNI	아오헝	인도네시아	10,000 미만
380	TUJ	투구틸	인도네시아	<10K
381	NPS	Nipsan	인도네시아	10,000 미만
382	UAN	Kuan	라오스	<10K
383	VBK	남서부 본톡	필리핀	10,000 미만
384	DMV	덤파스	말레이시아	10,000 미만
385	XKO	키오르	라오스	10,000 미만
386	KVE	칼라바칸 무루트	말레이시아	10,000 미만
387	MCM	말라카 포르투갈어 크리올어	말레이시아	<10K
388	LTU	Latu	인도네시아	<10K
389	GEF	게라이	인도네시아	<10K
390	CNC	Côông	베트남	10,000 미만
391	BPO	Anasi	인도네시아	10,000 미만
392	HLD	할랑 도안	라오스, 베트남	10,000 미만
393	NXK	코카크 나가	미얀마	<10K
394	PUJ	푸난 투부	인도네시아	<10K
395	XKN	카얀 강 카얀	인도네시아	<10K
396	YCP	체피아	라오스	10,000 미만
397	LCS	리사바타-누니알리	인도네시아	10,000 미만
398	HAF	하노이 수화	베트남	<10K
399	SLT	실라	라오스, 베트남	1만 명 미만

표 43: (5/6) 사용자가 1만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		동남아시아에 속하지 않음Crowd		
400	KVH	코모도	인도네시아	<10K
401	APF	파하난 아그타	필리핀	10,000 미만
402	BZB	안디오	인도네시아	10,000 미만
403	JAL	Yalahatan	인도네시아	<10K
404	MVR	마라우	인도네시아	<10K
405	AGZ	이리가 아그타 산	필리핀	10,000 미만
406	DKK	다카	인도네시아	10,000 미만
407	GAK	감코노라	인도네시아	<10K
408	KMD	마주카양 칼링가	필리핀	10,000 미만
409	MQP	마니파	인도네시아	<10K
410	PZN	제자라 나가	미얀마	10,000 미만
411	XKD	멘달람 카얀	인도네시아	<10K
412	XAY	카얀 마하캄	인도네시아	10,000 미만
413	XKY	우마 라산	인도네시아, 말레이시아	10,000 미만
414	MQQ	미노코크	말레이시아	10,000 미만
415	NEO	나메오	베트남	<10K
416	TLN	탈론도	인도네시아	<10K
417	BQY	카타 콜록	인도네시아	10,000 미만
418	MXR	무릭	말레이시아	10,000 미만
419	NTY	만치	베트남	10,000 미만
420	TEV	테오르	인도네시아	<10K
421	TTP	톰벨라라	인도네시아	<10K
422	AYT	마그부쿤 아이타	필리핀	10,000 미만
423	CKN	캉 친	미얀마	10,000 미만
424	CNO	Con	라오스	10,000 미만
425	GOQ	고라프	인도네시아	<10K
426	HOV	호보간	인도네시아	<10K
427	LPN	롱 푸리 나가	미얀마	10,000 미만
428	NLQ	라오 나가	미얀마	<10K
429	NQY	아캉 아리 나가	미얀마	10,000 미만
430	NUO	Ngoaun	라오스, 베트남	10,000 미만
431	PSG	페낭 수화	말레이시아	10,000 미만
432	UES	키오코	인도네시아	10,000명 미만

표 44: (6/6) 사용자가 1만 명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
동남아시아 인원				
1	SOW	소완다	인도네시아	<1K
2	DUV	Duvle	인도네시아	<1K
3	HMU	하맵	인도네시아	<1K
4	KTT	Ketum	인도네시아	<1K
5	MPZ	Mpi	태국	<1K
6	TVW	Sedoa	인도네시아	<1K
7	SYO	수옹	캄보디아	<1K
8	MGK	마웨스	인도네시아	<1K
9	MSS	웨스트 마셀라	인도네시아	<1K
10	DIJ	다이	인도네시아	<1K
11	DRN	서부 다마르	인도네시아	<1K
12	LJI	라이올로	인도네시아	<1K
13	MTH	Munggui	인도네시아	<1K
14	PSN	파나수안	인도네시아	<1K
15	RET	레타	인도네시아	<1K
16	TWG	Tereweng	인도네시아	<1K
17	BPG	봉고	인도네시아	<1K
18	AGT	중앙 카가얀 아그타	필리핀	<1K
19	KVZ	차우캄보	인도네시아	<1K
20	SKP	세카판	말레이시아	<1K
21	BSM	부사미	인도네시아	<1K
22	BZI	Bisu	태국	<1K
23	KZM	Kais	인도네시아	<1K
24	MHZ	모르	인도네시아	<1K
25	NKJ	나카이	인도네시아	<1K
26	PRU	Puragi	인도네시아	<1K
27	SKV	Skou	인도네시아	<1K
28	LAQ	카비아오	베트남	<1K
29	SSM	셈남	말레이시아	<1K
30	SLG	셀롱가이 무루트	인도네시아, 말레이시아	<1K
31	TPF	타르피아	인도네시아	<1K
32	VTO	Vitou	인도네시아	<1K
33	WSA	와렘보리	인도네시아	<1K
34	DGC	카시구란 두마갓 아그타	필리핀	<1K
35	BFE	베타프	인도네시아	<1K
36	KGB	카웨	인도네시아	<1K
37	KWH	코위아이	인도네시아	<1K
38	PPM	파푸마	인도네시아	<1K
39	TDI	토마디노	인도네시아	<1K
40	TMU	Iau	인도네시아	<1K
41	UKA	카부리	인도네시아	<1K
42	BKN	부키탄	인도네시아, 말레이시아	<1K
43	IMR	임로잉	인도네시아	<1K
44	TGQ	Tring	말레이시아	<1K
45	TLK	탈로키	인도네시아	<1K
46	ERT	에리타이	인도네시아	<1K
47	LPE	Lepki	인도네시아	<1K
48	VME	동 마셀라	인도네시아	<1K
49	MXZ	중앙 마셀라	인도네시아	<1K
50	AOS	타이캣	인도네시아	<1K
51	COG	Chong	태국	<1K
52	DPP	파파르	말레이시아	<1K
53	JET	마네	인도네시아	<1K
54	KAG	카자만	말레이시아	<1K
55	KGI	셀랑고르 수화	말레이시아	<1K
56	KLY	칼라오	인도네시아	<1K
57	KND	콘다	인도네시아	<1K
58	KUC	콴수	인도네시아	<1K
59	LVI	Lavi	라오스	<1K
60	NBN	쿠리	인도네시아	<1K
61	NER	야하디안	인도네시아	<1K
62	ONI	오닌	인도네시아	<1K
63	ORZ	Ormu	인도네시아	<1K
64	PKT	말랭	라오스, 베트남	<1K
65	RTH	라타한	인도네시아	<1K
66	SBT	킴키	인도네시아	<1K
67	TCM	타나메라	인도네시아	<1K
68	TRT	통가레	인도네시아	<1K
69	WTW	Wotu	인도네시아	<1K
70	XKQ	코로니	인도네시아	<1K
71	CWG	체크 웡	말레이시아	<1K
72	BPP	Kaure	인도네시아	<1K
73	ISD	Isnag	필리핀	<1K
74	PNA	푸난 바-비아우	말레이시아	<1K
75	SKZ	세카르	인도네시아	<1K
76	THM	아휴	태국	<1K
77	TOY	토포이요	인도네시아	<1K
78	DBE	Dabe	인도네시아	<1K
79	BVK	부캇	인도네시아	<1K
80	DEI	데미사	인도네시아	1,000명 미만

표 45: (1/3) 사용자가 1,000명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
동남아시아 인원				
81	JEL	Yelmek	인도네시아	<1K
82	NUN	아농	미얀마	<1K
83	OPK	코프카카	인도네시아	<1K
84	PAS	파파세나	인도네시아	<1K
85	TMJ	사마로케나	인도네시아	<1K
86	URN	우랑니린	인도네시아	<1K
87	XAU	카우웨라	인도네시아	<1K
88	KDY	케이자르	인도네시아	<1K
89	AUU	아우예	인도네시아	<1K
90	AUW	Awyi	인도네시아	<1K
91	FLH	Foau	인도네시아	<1K
92	GOP	Yeretuar	인도네시아	<1K
93	JAU	야우르	인도네시아	<1K
94	LHN	Lahanan	말레이시아	<1K
95	PEE	타제	인도네시아	<1K
96	PHQ	파나	라오스	<1K
97	TNZ	Ten'edn	말레이시아, 태국	<1K
98	WRU	와루	인도네시아	<1K
99	SVE	세필리	인도네시아	<1K
100	BGV	Warkay-Bipim	인도네시아	<1K
101	BHC	Biga	인도네시아	<1K
102	BQB	바구사	인도네시아	<1K
103	BSA	Abinomn	인도네시아	<1K
104	CCM	말라카 말레이 크리올어	말레이시아	<1K
105	GIQ	그린 겔라오	베트남	<1K
106	KJA	Mlap	인도네시아	<1K
107	KZV	Komyandaret	인도네시아	<1K
108	MRF	Elseng	인도네시아	<1K
109	SWR	Saweru	인도네시아	<1K
110	TAD	타우세	인도네시아	<1K
111	TBP	디브루드	인도네시아	<1K
112	TMO	Temoq	말레이시아	<1K
113	TYH	오두	라오스, 베트남	<1K
114	WUY	와우아이	인도네시아	<1K
115	XWR	크웨르바 맘베라모	인도네시아	<1K
116	RMH	Murkim	인도네시아	<1K
117	TML	탐닝 시탁	인도네시아	<1K
118	WET	페라이	인도네시아	<1K
119	BQQ	비리타이	인도네시아	<1K
120	BRS	바라스	인도네시아	<1K
121	BZU	Burmeso	인도네시아	<1K
122	EMW	엠폼라와스	인도네시아	<1K
123	KIQ	Kosare	인도네시아	<1K
124	KJY	키리키리	인도네시아	<1K
125	KNS	켄시우	말레이시아, 태국	<1K
126	LCC	Legenyem	인도네시아	<1K
127	MSO	Mombum	인도네시아	<1K
128	MVX	메오스와르	인도네시아	<1K
129	SAO	Sause	인도네시아	<1K
130	SNU	Viid	인도네시아	<1K
131	TLG	Tofanma	인도네시아	<1K
132	KGV	카라스	인도네시아	<1K
133	LNH	Lanoh	말레이시아	<1K
134	ASZ	As	인도네시아	<1K
135	KBI	Kaptiau	인도네시아	<1K
136	MSL	몰로프	인도네시아	<1K
137	WFG	조롬	인도네시아	<1K
138	DMU	테비	인도네시아	<1K
139	LLK	Lelak	말레이시아	<1K
140	TCQ	카이	인도네시아	<1K
141	AQN	노던 앨버타	필리핀	<1K
142	BNV	베네라프	인도네시아	<1K
143	ENC	En	베트남	<1K
144	ERW	에로크와나스	인도네시아	<1K
145	JBR	Jofotek-Bromnya	인도네시아	<1K
146	KHH	Kehu	인도네시아	<1K
147	KHP	카파우리	인도네시아	<1K
148	KXN	카노윗-탄중 멜라나우	말레이시아	<1K
149	MMB	모미나	인도네시아	<1K
150	NEC	네데방	인도네시아	<1K
151	NYL	뉴	태국	<1K
152	RAC	라사와	인도네시아	<1K
153	TNU	타이 캉	라오스	<1K
154	WOI	와레즈	인도네시아	<1K
155	YKI	Yoke	인도네시아	<1K
156	BED	Bedoanas	인도네시아	<1K
157	MZT	민팀	말레이시아	<1K
158	AGF	아르구니	인도네시아	<1K
159	APX	Aputai	인도네시아	<1K
160	KCD	Ngkâlmpw Kanum	인도네시아	1,000명 미만

표 46: (2/3) 사용자가 1,000명 미만인 동남아시아 원주민 언어.

번호	ISO 639-3	언어	지역	인구
		동남아시아 인원		
161	UGO	우공	태국	<1K
162	WBE	와리타이	인도네시아	<1K
163	MRA	Mlabri	라오스, 태국	<1K
164	AFZ	오보퀴타이	인도네시아	<1K
165	MGF	Maklew	인도네시아	<1K
166	TTN	Towei	인도네시아	<1K
167	KNQ	Kintaq	말레이시아	<1K
168	ULF	우스쿠	인도네시아	<1K
169	AWH	Awbono	인도네시아	<1K
170	BTI	부라테	인도네시아	<1K
171	BYL	바요노	인도네시아	<1K
172	DIY	Diuwe	인도네시아	<1K
173	KPI	Kofei	인도네시아	<1K
174	KRZ	소타 카눔	인도네시아	<1K
175	KWR	Kwer	인도네시아	<1K
176	TFO	테파로	인도네시아	<1K
177	TKX	탕코	인도네시아	<1K
178	TTI	토바티	인도네시아	<1K
		동남아시아에 없음Crowd		
179	LCD	Lola	인도네시아	<1K
180	ORS	오랑 셀레타르	말레이시아	<1K
181	KPD	코바	인도네시아	<1K
182	TRX	Tringgus-Sembaan Bidayuh	말레이시아	<1K
183	KQT	클리아스 강 카다잔	말레이시아	<1K
184	ATP	푸드톨 아타	필리핀	<1K
185	TCP	타우르 친	미얀마	<1K
186	KYD	Karey	인도네시아	<1K
187	PYY	Pyen	미얀마	<1K
188	TTW	룽 왓	말레이시아	<1K
189	XMX	살라와티	인도네시아	<1K
190	YMN	수눔	인도네시아	<1K
191	WKD	월	인도네시아	<1K
192	ABF	아바이 쑤가이	말레이시아	<1K
193	ESY	에스카얀	필리핀	<1K
194	KZB	카이보보	인도네시아	<1K
195	NJS	니사	인도네시아	<1K
196	NNI	북 누아울루	인도네시아	<1K
197	WHU	와하우 카얀	인도네시아	<1K
198	XKE	케레호	인도네시아	<1K
199	LCE	세카크	인도네시아	<1K
200	SDX	시부 멜라나우	말레이시아	<1K
201	BFK	반 코어 수화	태국	<1K
202	KAX	카오	인도네시아	<1K
203	SRK	세루둥 무루트	말레이시아	<1K
204	PUD	푸난 아풋	인도네시아	<1K
205	BGY	벵고이	인도네시아	<1K
206	KZD	카다이	인도네시아	<1K
207	KVP	콤파네	인도네시아	<1K
208	AUQ	아누스	인도네시아	<1K
209	AZT	Faire Atta	필리핀	<1K
210	HUD	Huaulu	인도네시아	<1K
211	LGH	라구	베트남	<1K
212	TIP	Trimuris	인도네시아	<1K
213	TYJ	타이요	라오스, 베트남	<1K
214	TYS	타이 사파	베트남	<1K
215	MQI	마리리	인도네시아	<1K
216	PDN	페단	인도네시아	<1K
217	MNQ	Minriq	말레이시아	<1K
218	DAZ	다오	인도네시아	<1K
219	GNQ	가나	말레이시아	<1K
220	LRN	로랑	인도네시아	<1K
221	BSU	Bahonsuai	인도네시아	<1K
222	PUC	푸난 메라프	인도네시아	<1K
223	RMX	로맘	베트남	<1K
224	TYL	Thu Lao	베트남	<1K
225	YRS	야르순	인도네시아	<1K
226	ATL	이라야 아그타 산	필리핀	<1K
227	PUF	푸난 메라	인도네시아	<1K
228	UMI	우킷	말레이시아	<1K
229	JVD	Javindo	인도네시아	<1K
230	SRT	사우리	인도네시아	1,000명 미만

표 47: (3/3) 사용자가 1,000명 미만인 동남아시아 토착어.

번호	ISO 639-3	언어	지역	인구
		SEACrowd에서		
1	MNU	Mer	인도네시아	<100
2	ITX	이틱	인도네시아	<100
3	KXQ	스마르키 카눔	인도네시아	<100
4	LIX	리야부쿠	인도네시아	<100
5	AWR	아웨라	인도네시아	<100
6	BDX	부동-부동	인도네시아	<100
7	IRE	에레시아	인도네시아	<100
8	TDS	두타이	인도네시아	<100
9	MRX	Dineor	인도네시아	<100
10	AMQ	아마하이	인도네시아	<100
11	KZU	카유풀라우	인도네시아	<100
12	MOK	모로리	인도네시아	<100
13	PLH	파울로히	인도네시아	<100
14	SGU	살라스	인도네시아	<100
15	AIP	부루마쿡	인도네시아	<100
16	DBN	두리안케레	인도네시아	<100
17	DUL	이나그타 알라바트	필리핀	<100
18	MOQ	모르	인도네시아	<100
19	NAA	남랴	인도네시아	<100
20	MVS	Massep	인도네시아	<100
21	AEM	아렘	라오스, 베트남	<100
22	MQR	만데르	인도네시아	<100
23	XKW	켐브라	인도네시아	<100
24	KKB	크웨리사	인도네시아	<100
25	ATZ	아르타	필리핀	<100
26	IBH	Bih	베트남	<100
27	KHD	바디 카눔	인도네시아	<100
28	NUL	누사 라우트	인도네시아	<100
29	SCQ	Chung	캄보디아	<100
30	MQT	목	미얀마, 태국	<10
31	BTJ	바카네즈 말레이	인도네시아	<10
32	WOR	워리아	인도네시아	<10
33	SPI	사포니	인도네시아	<10
34	DSN	Dusner	인도네시아	<10
35	LGI	Lengilu	인도네시아	<10
36	BTN	라타뇽	필리핀	<10
37	TNI	탄디아	인도네시아	<10
38	HUW	후쿠미나	인도네시아	<10
39	KZL	카엘리	인도네시아	<10
40	SXM	삼레	캄보디아, 태국	<10
41	HPO	Hpon	미얀마	<10
42	MPY	Mapia	인도네시아	<10
43	NIL	Nila	인도네시아	<10
44	SBO	사봄	말레이시아	<10
45	SRW	세루아	인도네시아	<10
46	TAS	타이 보이	베트남	<10
47	XBN	케나보이	말레이시아	<10
48	XXT	탐보라	인도네시아	<10
		동남아시아에 속하지 않음		
49	ORN	오랑 카나크	말레이시아	<100
50	LVA	마쿠바	동티모르	<100
51	SPG	Sihan	말레이시아	<100
52	IBU	이부	인도네시아	<100
53	PNM	푸난 바투	말레이시아	<100
54	CSD	치앙마이 수화	태국	<100
55	AYS	소르소곤 아이타	필리핀	<100
56	LIO	리키	인도네시아	<100
57	PEY	Petjo	인도네시아	<100
58	HTI	호티	인도네시아	<100
59	HUK	홀룽	인도네시아	<100
60	ISM	마시마시	인도네시아	<100
61	KZX	카마리안	인도네시아	<100
62	PNS	포노사칸	인도네시아	<100
63	AGK	카투붕 아그타	필리핀	<10
64	NAE	나카엘라	인도네시아	<10
65	ATM	아타	필리핀	<10
66	IHB	이하 기반 피진어	인도네시아	<10
67	TVY	티모르 피진어	동티모르	<10
68	DUY	디카마이 아그타	필리핀	<10
69	DYG	빌라 비시오사 아그타	필리핀	<10
70	LOX	Loun	인도네시아	<10
71	ONX	오닌 기반 피진어	인도네시아	<10
72	TCL	타만	미얀마	<10
73	VMS	모크셀라	인도네시아	<10
74	WEA	웨와우	미얀마	<10

표 48: SEA 지역에서 100명 미만의 사용자를 가진 토착어.