Direct Cross-Language Search for ASEAN: Connecting ASEAN Regional Languages with Standard Burmese and Ethnic Languages

Hsu Myat Mo

University of Computer Studies, Yangon (UCSY), Myanmar

Background : Language Barrier in ASEAN

- *The Digital Divide*: ASEAN integration is hindered by language barriers. Current Cross-Lingual Information Retrieval (CLIR) systems rely on English as a "pivot" (Language A → English → Language B), causing translation errors, cultural loss, and high latency.

- *The Representation Gap*: Existing multilingual AI models heavily favor major national languages (Thai, Vietnamese, Indonesian) but severely underrepresent Myanmar's 100+ ethnic languages (e.g., Mon, Rakhine, Myeik, Shan).

- *The Consequence*: Millions of ethnic language speakers are excluded from the digital ASEAN ecosystem, lacking access to critical cross-border information.

Project Targets

- **Target 1:**
  - ***Direct Connection*:** Build a CLIR system that bypasses the English pivot, connecting ASEAN languages directly (e.g., Shan ↔ Thai, Mon ↔ Burmese).
- **Target 2:**
  - ***Inclusive AI:*** Integrate Myanmar's ethnic languages into the regional AI framework, ensuring no community is left behind.
- **Target 3:**
  - ***Practical Application*:** Apply the system to a high-impact, non-sensitive domain (Agriculture / Disaster Resilience) to directly benefit local communities.

Proposed Method : Core Model & Hybrid Retrieval Architecture

- ***Base Models*:**
  - State-of-the-art multilingual embedding model such as BGE-M3 with native support for hybrid retrieval.
  - Handles long-context documents, ideal for document-level ethnic language data.

- ***Hybrid Retrieval Architecture.***
  - *Dense Search (Semantic):* Captures deep meaning across languages (e.g., Mon ↔ Burmese).
  - *Sparse Search (Lexical):* Acts as a safety net by matching exact keywords, cognates, and domain-specific terminology (e.g., shared agricultural terms between Shan and Thai).

- ***Why Hybrid Matters for Low-Resource Data***
  - Real-world ethnic data is noisy and misaligned — hybrid search ensures robustness even when semantic embeddings are imperfect.

Proposed Method : Pipeline & Training Strategy

- ***Data Generation:***
  - Generate synthetic parallel corpora between standard Burmese and ethnic languages (e.g., Mon, Rakhine, Myeik, Shan).
- ***Retrieval:***
  - Fine-tune via contrastive learning on the generated parallel data to map ethnic languages directly to Burmese and ASEAN neighbors
- ***Re-Ranking***
  - Deploy SEA-LION as a cross-encoder to re-rank top-k retrieved documents.
- ***Parameter-Efficient Fine-Tuning (PEFT)***
  - Apply LoRA (Low-Rank Adaptation) to extend model capabilities to low-resource ethnic languages without "forgetting" major ASEAN languages.

Implementation: Handling Real-World Low-Resource Data

- ***The Data Challenge*:** Real-world ethnic data is multi-domain, noisy, and often misaligned at the sentence level.

- **Our Solution:**
  - **Document-Level Alignment:** Match whole paragraphs/documents instead of forcing fragile sentence-level pairs.
  - **Two-Stage Training:**
    - *Stage 1:* Continual pre-training on all multi-domain data to learn vocabulary and grammar.
    - *Stage 2:* Supervised fine-tuning on a specific domain (e.g., Agriculture) for the retrieval task.

- **Synthetic Queries:** Use LLMs to generate diverse search queries from documents, validated by native speakers, overcoming data scarcity.

Experiments and Field Testing

- **Benchmark Creation:** Develop the "ASEAN-CLIR" benchmark, the first to include Myanmar ethnic languages alongside major ASEAN languages.

- **Evaluation Metrics:** Measure retrieval performance using standard Information Retrieval metrics:
  - Recall@K
  - Mean Reciprocal Rank (MRR)
  - Normalized Discounted Cumulative Gain (NDCG)

- **Field Testing:** Evaluate the system's practical utility in retrieving agricultural or disaster safety information across language barriers.

Impact : Scientific & Technological

- ***Bridging the AI Gap*:** Proves that severely low-resource, underrepresented ethnic languages can be successfully integrated into large-scale regional AI systems.
- ***Methodology*:** Establishes a robust framework for handling noisy, multi-domain, and misaligned low-resource data using document-level alignment and two-stage training.
- ***Open Benchmark*:** Provides the first CLIR evaluation benchmark for intra-ASEAN ethnic language pairs, filling a critical gap in NLP research.

Impact: Societal & Collaborative

- **Societal Impact:**
  - ***Digital Inclusion:*** Empowers millions of ethnic language speakers to access critical digital information in their native tongues.
  - ***Cultural Preservation:*** Validates and digitizes ethnic languages in the AI era, preventing digital marginalization.

- **Collaborative Impact:**
  - Fosters strong research ties between Myanmar (UCSY) and ASEAN partners

Output/Outcome: Scientific & Societal

- ***Scientific Outputs*:**
  - Peer-reviewed publications at top-tier AI/NLP conferences (e.g., ACL, EMNLP, COLING).
  - Open-source, fine-tuned retrieval model weights optimized for Southeast Asian ethnic languages.

- ***Societal Outputs*:**
  - A public, open-source "ASEAN-CLIR" dataset available for the global research community.
  - A practical, deployable search tool/API for local NGOs, agricultural extension workers, and disaster response teams.

Output/Outcome: Collaborative

- ***New Partnerships*:** Establishes a foundational collaboration network between Myanmar researchers and ASEAN IVO member institutions.
- ***Community Building*:** Launches a sustainable "ASEAN Language AI" open-source ecosystem, inviting continuous contributions from regional universities.
- ***Future Readiness*:** Creates the groundwork for future multimodal (text + image) cross-lingual search capabilities.

Conclusion

- ***Targets*:** Designed a direct, English-free CLIR system that specifically includes Myanmar's marginalized ethnic languages alongside major ASEAN languages.
- ***Method*:** Utilizes a robust training approach and synthetic query generation to handle real-world, noisy, low-resource data.
- ***Scientific Impact*:** Delivers a novel intra-ASEAN CLIR benchmark and proves the viability of integrating severely low-resource languages into regional AI.
- ***Societal Impact*:** Directly supports ASEAN's digital inclusion goals by providing local communities with native-language access to vital agricultural and disaster resilience information.

References:

- **Chen, J., et al. (2024).** *BGE-M3: Multilingual, Multi-Granularity, and Multi-Function Text Embedding Models.* arXiv preprint.
- **Costa-jussà, M. R., et al. (2022).** *No Language Left Behind: Scaling Human-Centered Machine Translation.* arXiv preprint arXiv:2207.04672.
- **Tjhi, W. C., et al. (2024).** *SEA-LION: Southeast Asian Languages in One Network.* AI Singapore.
- **Aung, T., et al. (2024).** *Burmese-SAN: Burmese NLP Benchmark for Evaluating Large Language Models*.
- **Zhang, X., et al. (2022).** *MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages.* Proceedings of EMNLP.

Thank You!