Cross-Lingual Fact-Check Retrieval Using Contrastive Learning and Knowledge Distillation

Web Retrieval and Mining Final Project Proposal

Yi-Xuan JIANG   James   Ting-Kuan HSIEH

National Taiwan University

May 2025

Outline

Problem Statement

- **Challenge:** Misinformation spreads rapidly across languages
- **Task:** SemEval-2025 Task 7 - Multilingual and Crosslingual Fact-Check Retrieval
- **Goal:** Retrieve relevant fact-checked claims for social media posts across multiple languages

Why This Matters

1. **Resource Optimization** for fact-checking organizations
2. **Cross-Border Information Flow** tracking
3. **Timely Response** to viral misinformation
4. **Broader Coverage** across language barriers
5. **Societal Impact** on public discourse and decision-making

SemEval 2025 Task 7 Overview

- **Task:** Given a social media post, retrieve relevant fact-checked claims from a large multilingual database
- **Languages:** English, Spanish, French, German, Italian, Portuguese, Arabic, Hindi, etc.
- **Primary Challenge:** Finding semantic equivalence across languages
- **Data:** MultiClaim dataset with posts and fact-checks in multiple languages
- **Evaluation:** success@k metrics (primarily success@10)

Multilingual Dense Retrieval Architecture

- Dual-encoder architecture using multilingual language models
- Encode posts and fact-checks into a shared embedding space
- Base models: XLM-RoBERTa, mBERT, LaBSE

Key Technical Approaches

- **Contrastive Learning**
  - In-batch negatives and hard negative mining
  - Training with positive and negative examples
- **Knowledge Distillation**
  - Using larger teacher models (multilingual-T5, BLOOM)
  - Distilling knowledge into efficient student models
- **Translation-Augmented Training**
  - Data augmentation with translations
  - Learning semantic equivalence across languages

Retrieval Enhancement Techniques

Query Expansion

- Multilingual WordNet/BabelNet
- Adding synonyms and related terms

Reranking

- Two-stage retrieval approach
- Lightweight initial retrieval (BM25)
- Cross-encoder reranking

Multimodal Integration

- OCR for text in images
- Combining image text with post text

Language-Specific Processing

- Language identification
- Custom tokenization and normalization

Generative Pseudo Labeling for Unsupervised Multilingual Retrieval

- Approach from Chen et al. (EACL 2024)
- Creating synthetic training data for low-resource languages
- Using LLMs to generate pseudo-aligned text pairs
- Helping bridge gaps between languages with limited parallel data

Anticipated Challenges

- **Semantic Drift Across Languages**
  - Concepts may not map perfectly between languages
- **Computational Efficiency**
  - Balancing performance with inference speed
- **Data Imbalance**
  - Varying amounts of training data across languages
- **Cultural Context**
  - Misinformation often relies on cultural references
- **Translation Quality**
  - Variable quality of translations in the dataset

Evaluation Metrics

- **Primary Metric** (SemEval Task):
  - success@10 - Whether the correct fact-check appears in top 10 results
- **Additional Internal Metrics**:
  - Mean Reciprocal Rank (MRR)
  - Normalized Discounted Cumulative Gain (NDCG)
  - Precision and Recall at various cutoffs (P@k, R@k)
  - Language-specific performance analysis

Ablation Studies

Component Contribution Analysis:

- Performance with/without OCR text
- Comparing different embedding models
- Impact of knowledge distillation
- Value of translation augmentation
- Effect of generative pseudo labeling

Language-Specific Analysis:

- Performance across different language pairs
- Impact of language families on crosslingual retrieval

Datasets and Resources

- **Primary Dataset**:
  - MultiClaim dataset from SemEval-2025 Task 7
- **Additional Resources**:
  - CLEF CheckThat! Lab datasets
  - MultiLingual Misinformation Dataset (MLMD)
  - Fact-checking websites (Snopes, PolitiFact, etc.)
- **Technical Stack**:
  - Hugging Face Transformers
  - PyTorch
  - FAISS for efficient similarity search

Experimental Setup

- **Cross-Validation Approach**:
  - 5-fold CV preserving language distribution
  - 4 folds for training, 1 for validation
- **Hyperparameter Tuning**:
  - Bayesian optimization
  - Focus on embedding size, learning rate, batch size
- **Analysis**:
  - Language-specific performance evaluation
  - Error analysis by claim type and language pair

Project Timeline

Date	Milestone	Activities
May 2	Proposal submission	Submit proposal document
May 3-8	Data exploration	Preprocessing, repository setup
May 9	Feedback review	Adjust plans based on feedback
May 10-16	Baseline models	BM25, TF-IDF implementation
May 17-23	Core development	Multilingual embedding models
May 24-30	Advanced features	Contrastive learning, distillation
May 31-Jun 4	Finalization	System integration, testing
Jun 6	Final submission	Presentation and report

References

1. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019*.
2. Conneau, A., et al. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *ACL 2020*.
3. Feng, F., et al. (2022). Language-agnostic BERT Sentence Embedding. *ACL 2022*.
4. Nakov, P., et al. (2021). The CLEF-2021 CheckThat! Lab on Detecting Check-Worthy Claims, Previously Fact-Checked Claims, and Fake News. *ECIR 2021*.
5. Shaar, S., et al. (2020). That is a Known Lie: Detecting Previously Fact-Checked Claims. *ACL 2020*.
6. **Chen, P., et al. (2024). Unsupervised Multilingual Dense Retrieval via Generative Pseudo Labeling. EACL 2024.**

Thank You!

Questions?

Yi-Xuan JIANG: b10902010@ntu.edu.tw
James: r13921031@ntu.edu.tw
Ting-Kuan HSIEH: b10701166@ntu.edu.tw