English-Chinese Cross-lingual News Retrieval System

Lingxiao Zhong, Yixuan Chen, Anlin Ma

Keywords: Word Embedding, Cross-lingual Mapping, Ranking.

1 Introduction

Access to information across language barriers remains a significant challenge in our globalized world. International businesses tracking foreign markets, journalists researching global incidents, and policy analysts monitoring worldwide developments need to find relevant news across languages. Although English dominates online news, crucial perspectives and breaking stories often appear first in regional languages such as Chinese.

This project develops a cross-lingual information retrieval system that allows users to query in one language and retrieve relevant news articles from both English and Chinese sources. Leveraging word embeddings, cross-lingual alignment techniques, and retrieval-augmented generation (RAG), we create a unified search experience across both language ecosystems. While recent pre-trained multilingual models achieve strong accuracy, they require substantial computational resources. Our approach aims to balance retrieval quality with efficiency, making it more suitable for resource-constrained deployment scenarios. Success would benefit international businesses, academic researchers, and news organizations while demonstrating a scalable approach extensible to other language pairs and domains.

2 IR Task Definition

2.1 Task Overview

Given a query in one language (source language), retrieve the most relevant news articles from both source and target language news dataset, presenting results in both languages with optional contextualized summaries.

2.2 Specific Problems

Our system will handle bilingual retrieval regardless of query language

- English query → Relevant articles from both English and Chinese sources
- Chinese query → Relevant articles from both English and Chinese sources

2.3 Example Queries

- **Query Type:** Natural language questions or keyword queries
- **Query Language:** English or Chinese
- **Query Length:** 5-50 words typical range
- **Query Examples:**
  - **English:** “Tesla factory opening in Shanghai”
  - **English:** “What are China’s recent AI regulations?”
  - **Chinese:** “美国通货膨胀率” (US inflation rate)
  - **Chinese:** “乌克兰战争最新进展” (Latest developments in Ukraine war)

2.4 Example Documents

- **Document Type:** News articles from major English and Chinese news sources

- **Languages:** English articles (CNN, Reuters, BBC, NYT) and Chinese articles (新华社/Xinhua, 人民日报/People's Daily, 财新/Caixin)
- **Document Fields:** Title, body text, publication date, source
- **Time Range:** Articles from 2023-2025

2.5 Output Specifications

- **Primary Output:** Ranked list of top-k (k=10) relevant news articles from both English and Chinese sources, interleaved by relevance score
- **Secondary Output (RAG component):** Generated summary or answer in query language synthesizing information from retrieved articles across both languages
- **Output Format:** Each retrieved article includes title, snippet, relevance score, source, and language indicator

2.6 Implicit Information Need

The system addresses the need to discover comprehensive information about events, topics, or entities by:

1. Aggregating perspectives from both English and Chinese sources
2. Overcoming language barriers without requiring user fluency in both languages
3. Providing timely access to breaking news regardless of publication language
4. Enabling comprehensive coverage through multilingual retrieval

3 Data

3.1 Primary Data Source

HC4: A New Suite of Test Collections for Ad Hoc CLIR [8]

- Monolingual documents in target language (Chinese, Persian, Russian), drawn from the Common Crawl News corpus.
- Queries in English as well as translated version in target language.
- Used an active learning system (HiCAL) to give relevance judgments.
- **Size**: 0.65M docs (train), 0.49M (eval); 60 queries (10 train, 50 eval) for Chinese.

3.2 Backup Data Source

CLIRMatrix: A massively large collection of bilingual and multilingual datasets for Cross-Lingual Information Retrieval [12]

- 50.1 million documents from Wikipedia across 139 languages.
- A bilingual dataset of queries in one language matched with relevant documents in another language for 139x138=19,182 language pairs, with up to 10,000 queries per language pair for train and up to 1,000 for val/test.
- Relevance score generated from monolingual IR based on BM25.

3.3 Bilingual Dictionary

- **Source:** MUSE bilingual dictionaries (Facebook Research) [3], CC-CEDICT [10]
- **Size:** 21597 in MUSE, 124026 in CC-CEDICT
- **Purpose:** Anchor points for cross-lingual embedding alignment
- **Format:** Word pairs with translation equivalences

4 Related Work

Cross-Language Information Retrieval (CLIR) aims to retrieve documents in one language using queries written in another. The chief challenge—beyond monolingual IR—is translating lexical meaning well enough to preserve intent while coping with ambiguity, morphology, and domain shift.

Early CLIR systems (1990s–2010s) primarily relied on dictionary-based or machine translation (MT)-based query translation [1, 11]. These approaches translated either the query, the documents, or both, before applying traditional retrieval models such as TF-IDF or BM25. Dictionary-based systems suffered from ambiguity and limited coverage, while MT-based methods improved precision but introduced translation noise and required significant computational resources. Some hybrid approaches used parallel corpora or bilingual lexicons to build statistical translation models [6], paving the way for learning-based CLIR methods.

With the emergence of neural networks and large language models (LLMs) since 2018, modern CLIR shifts from explicit translation to shared semantic spaces and neural ranking. Large pre-trained multilingual language model were trained to produce strong cross-lingual sentence embeddings widely used as CLIR bi-encoder backbones or for hybrid retrieval [5]. In addition, with improved machine translation results thanks to LLMs, translate-then-search remains a solid baseline. Results show that high-quality neural MT + neural IR is a very strong baseline but increases indexing/serving costs; late-interaction multilingual models can match quality with lower translation overhead [7]. The development of LLMs also allows for large-scale synthetic/weakly-parallel data. Synthetic CLIR training sets (e.g., LLM/MT-generated) and massive weakly parallel corpora (e.g., Wikipedia alignments) help scale supervision to low-resource pairs [9].

5 Evaluations and Results

5.1 Evaluation Metrics

Primary Metrics

1. **Precision@k (P@k)** - Proportion of relevant documents in top-k results
   - We will report P@5 and P@10
   - Directly measures user experience (most users only check first page)
2. **Mean Average Precision (MAP)** - Average of precision scores at each relevant document position
   - Rewards systems that rank relevant documents higher
   - Standard metric for ranked retrieval evaluation
3. **Normalized Discounted Cumulative Gain (NDCG@k)** - Considers graded relevance and position
   - Uses our 3-point relevance scale effectively
   - NDCG@10 will be our primary metric for comparing systems

Efficiency Metrics:

Given that pre-trained multilingual models like XLM-RoBERTa [2] and mBERT [4] exist, we will evaluate computational efficiency to demonstrate the practical advantages of our lightweight embedding approach:

1. **Query Latency** - Average time to process query and retrieve top-10 results
   - Measured in milliseconds
   - Important for real-time applications
   - Expected advantage: Custom embeddings should be 10-50x faster than transformer models
2. **Memory Footprint** - Peak RAM usage during retrieval
   - Measured in GB

    - Critical for deployment on resource-constrained environments
3. **Index Build Time** - Time to encode entire document collection
    - One-time cost but important for scalability
4. **Throughput** - Queries processed per second
    - Measures system scalability
    - Important for production deployment with multiple concurrent users

5.2 Baseline Systems

- Random retriever: Retrieve k documents randomly.
- Translate-then-search: Machine translate the query to target language, then run BM25 with no hyperparameter tuning.

5.3 Expected Results

We expect our method to perform better than the translate-then-search baseline but worse than state-of-the-art systems that are more complex or involve pre-trained large multilingual models. We also expect our system of utilize less resources and run faster than the those complex systems or pre-trained models.

6 Work Plan

- Week 7: Finalize proposal.
- Week 8: Data collection.
- Week 9: Implement baselines and check validity of the data.
- Week 10-11: Implement our planned methodology.
- Milestone 1: Write an initial draft report with data, baselines, and progress of our methodology for the first project update.
- Week 12-13: Evaluation and improvements.
- Week 14: Prepare presentation and final report.
- Milestone 2: Present project status in class.
- Week 15-16: Extra time plan in case any of the steps take longer than expected.

References

[1] Lisa Ballesteros and W. Bruce Croft. Dictionary-based methods for cross-language information retrieval. *Information Processing & Management*, 36(3):353–376, 2000.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. URL: `https://aclanthology.org/2020.acl-main.747/`, `doi:10.18653/v1/2020.acl-main.747`.

[3] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017. URL: `https://github.com/facebookresearch/MUSE?tab=readme-ov-file`.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill

Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL: `https://aclanthology.org/N19-1423/`, `doi:10.18653/v1/N19-1423`.

[5] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL: `https://aclanthology.org/2022.acl-long.62/`, `doi:10.18653/v1/2022.acl-long.62`.

[6] Victor Lavrenko, Mathieu Choquette, and W Bruce Croft. Cross-lingual relevance models. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 175–182, 2002.

[7] Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. Overview of the trec 2022 neuclir track, 2023. URL: `https://arxiv.org/abs/2304.12367`, `arXiv:2304.12367`.

[8] Dawn Lawrie, James Mayfield, Douglas W. Oard, and Eugene Yang. Hc4: A new suite of test collections for ad hoc clir. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 351–366, Cham, 2022. Springer International Publishing.

[9] James Mayfield, Eugene Yang, Dawn Lawrie, Samuel Barham, Orion Weller, Marc Mason, Suraj Nair, and Scott Miller. Synthetic cross-language information retrieval training data, 2023. URL: `https://arxiv.org/abs/2305.00331`, `arXiv:2305.00331`.

[10] MDBG Chinese-English Dictionary. CC-CEDICT: A chinese-english dictionary, 2025. Creative Commons Attribution-ShareAlike 4.0 License. URL: `https://www.mdbg.net/chinese/dictionary?page=cc-cedict`.

[11] Jian-Yun Nie. *Cross-Language Information Retrieval*. Morgan & Claypool Publishers, 2010.

[12] Shuo Sun and Kevin Duh. CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online, November 2020. Association for Computational Linguistics. URL: `https://aclanthology.org/2020.emnlp-main.340/`, `doi:10.18653/v1/2020.emnlp-main.340`.