Neural Methods for Cross-Language Information Retrieval

Dawn Lawrie
HLTCOE, Johns Hopkins University
Baltimore, MD, USA
lawrie@jhu.edu

James Mayfield
HLTCOE, Johns Hopkins University
Baltimore, MD, USA
mayfield@jhu.edu

Suraj Nair
University of Maryland
College Park, MD, USA
srnair@umd.edu

Douglas W. Oard
University of Maryland
College Park, MD, USA
oard@umd.edu

Eugene Yang
HLTCOE, Johns Hopkins University
Baltimore, MD, USA
eugene.yang@jhu.edu

ABSTRACT

This half day tutorial introduces the participant to the basic concepts underlying neural Cross-Language Information Retrieval (CLIR). It discusses the most common algorithmic approaches to CLIR, focusing on modern neural methods; the history of CLIR; where to find and how to use CLIR training collections, test collections and baseline systems; how CLIR training and test collections are constructed; and open research questions in CLIR.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Evaluation of retrieval results*; *Test collections*; **Multilingual and cross-lingual retrieval**; *Relevance assessment*; *Similarity measures*; Probabilistic retrieval models; Combination, fusion and federated search; Query representation; *Retrieval effectiveness*; *Retrieval efficiency*.

KEYWORDS

Cross-language information retrieval, CLIR, Tutorial

ACM Reference Format:
Dawn Lawrie, James Mayfield, Suraj Nair, Douglas W. Oard, and Eugene Yang. 2023. Neural Methods for Cross-Language Information Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3539618.3594244

1 MOTIVATION

Cross-Language Information Retrieval (CLIR) is the identification of documents in one language that are relevant to an information need that was expressed in a different language. CLIR has been studied for decades, and efficient non-neural algorithms are available that are reasonably effective. Yet until recently, the impact of CLIR has been muted in commercial practice, perhaps because finding a relevant document written in a different language begged the question of how that document could be used. While there are use cases in which the searcher has the language skills to read the retrieved documents, or could afford human translation of those documents, there are many more use cases in which they do not.

The advent of transformer models and large language models has changed this landscape dramatically in two ways. First, high-quality machine translation (MT) between many languages is now possible. When accurate and fluent MT from the language of the documents to the language of the queries is available, the potential utility of CLIR jumps dramatically. Second, CLIR techniques that exploit contextual embedding models have resulted in retrieval effectiveness that is qualitatively improved over that of the best statistical CLIR models.

Holding a CLIR tutorial at SIGIR at this time has several motivations. First, the problem is well circumscribed, allowing a tutorial to quickly focus on relevant technologies and metrics. Second, CLIR resources are now widely available, and there is an active CLIR research community, served by the TREC NeuCLIR track, which allows researchers to impartially evaluate their work. Third, neural methods—and particularly transformer-based methods such as BERT—have only been widely applied to Information Retrieval (IR) in the last few years, with CLIR approaches lagging somewhat. Thus, there is currently much room for innovation in neural CLIR architectures. Fourth, generating the very large collections of CLIR training data needed for fine-tuning neural methods to a retrieval task has until recently relied heavily on machine translation of monolingual resources such as MS MARCO [1]. However, the large generative language models now coming online suggest new ways to create CLIR training data that were not previously possible, opening new possibilities for targeting such training data directly at domains or genres of interest.

CLIR tutorials on dictionary-based and statistical techniques were common around the turn of the century, including at SIGIR, most recently in 2000 and 2007. More recently, statistical techniques for CLIR have been included in tutorials on other topics (e.g., the SIGIR Wikipedia tutorial in 2015). However, all of these predate the neural revolution in information retrieval that resulted from the introduction of transformer models. Moreover, the most recent widely-cited CLIR tutorial (Nie 2010, [4]) also predates the neural revolution, and of the five CLIR survey papers published since of which we are aware, only one [2] includes any focus on neural CLIR methods. As the focus on neural techniques has expanded there have also been dedicated tutorials on multilingual embedding models in other venues, including for example at EMNLP 2017; but even that specialized topic has not yet been the principal focus of

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3594244

any SIGIR tutorial. Thus, we now offer a timely tutorial at SIGIR that focuses on neural methods for CLIR.

2 OBJECTIVES

Upon completion of the half-day tutorial, participants will have a basic working understanding of:

- the most common algorithmic approaches to CLIR, focusing on modern neural methods;
- the history of CLIR;
- where to find and how to use CLIR training collections, CLIR test collections, and CLIR baseline systems;
- how CLIR training and test collections are constructed; and
- open research questions in CLIR.

3 INTENDED AUDIENCE

The tutorial participant should have a basic understanding of monolingual text retrieval approaches and evaluation. Basic knowledge of IR (such as BM25 ranking and TREC-style evaluations) is presumed, but prior exposure to CLIR is not. Coding skills are not required, although the tutorial includes a practicum with sample code for those interested. The tutorial informs two levels of student: the novice with little exposure to CLIR, and the researcher with some exposure to CLIR and some understanding of current monolingual neural methods. All of the basic concepts are introduced briefly; and the tutorial addresses the main topics listed below in more technical detail.

4 FORMAT AND TOPICS

The tutorial is presented as a set of lectures followed by a brief practicum. It is designed for a hybrid audience consisting of both in-person and remote participants.

After a brief introduction, the tutorial dives into neural methods, using as examples a limited set of test collections. The second half of the tutorial revisits issues we glossed over during the first half, including examples of text processing for different languages, test collections for CLIR and for Multilingual Information Retrieval (MLIR), and training data. The tutorial focuses on text retrieval throughout, although of course many of the ideas can be extended to cross-language retrieval from other modalities such as spoken content or printed materials.

The following is an outline of the tutorial:

- Introduction
  - The CLIR task
  - Goals of the tutorial
  - A brief history of CLIR
  - Crossing the language barrier
- Neural CLIR Algorithms
  - Effective CLIR: Cross-encoders
  - Efficient CLIR: Bi-encoders
  - Efficient CLIR: Sparsification
  - Effective CLIR: Reranking
  - Effective CLIR: Ensembling
- Multilingual Text Processing
  - Encodings, writing systems, morphology and tokenization
- CLIR Evaluation Data
  - Requirements
  - Extant evaluation data
- CLIR Training Data
  - Requirements
  - Extant training data
  - Training data generation
- Multilingual Retrieval
  - Evaluation
  - Multilingual algorithms
  - Machine translation and run merging
- Practicum (continued asynchronously online)

The practicum is designed to introduce attendees to practical aspects of CLIR. The tutorial includes Jupyter (Google Colab) notebooks covering training, indexing, retrieval, and evaluation of CLIR systems over a CLIR evaluation collection available on `ir_datasets` [3]. We provide simple code examples for both sparse and neural retrieval methods.

5 SUPPORTING MATERIALS

Attendees receive the following materials, all of which we expect to share in persistent forms for future use by others:

- a copy of the tutorial slides;
- a pointer to a video of the tutorial;
- a reading list, aligned to tutorial sections; and
- access to a GitHub repo that includes sample code, configuration files, Jupyter notebooks for the practicum, and a set of relevant links.

These materials will be available at https://github.com/hltcoe/clir-tutorial.

REFERENCES

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268* (2016).

[2] Petra Galuščáková, Douglas W Oard, and Suraj Nair. 2021. Cross-language information retrieval. *arXiv preprint arXiv:2111.05988* (2021).

[3] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified Data Wrangling with ir_datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR).*

[4] Jian-Yun Nie. 2010. *Cross-Language Information Retrieval.* Morgan & Claypool Publishers. https://doi.org/10.2200/S00266ED1V01Y201005HLT008