MODELING CROSS-LINGUAL KNOWLEDGE IN MULTILINGUAL INFORMATION RETRIEVAL SYSTEMS

A Dissertation Presented

by

ZHIQI HUANG

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2024

Robert and Donna Manning College of
Information and Computer Sciences

MODELING CROSS-LINGUAL KNOWLEDGE IN MULTILINGUAL INFORMATION RETRIEVAL SYSTEMS

A Dissertation Presented

by

ZHIQI HUANG

Approved as to style and content by:

James Allan, Chair

Razieh Rahimi, Member

Mohit Iyyer, Member

Jeffrey Dalton, Member

Ramesh K. Sitaraman, Associate Dean for
Educational Programs and Teaching
Robert and Donna Manning College of
Information and Computer Sciences

To my grandparents.

ACKNOWLEDGMENTS

I would like to express my gratitude to those who have supported and guided me throughout my Ph.D. study.

First and foremost, I would like to thank my adviser, James Allan, for his thoughtful consideration and steadfast support. His guidance consistently deepens my understanding of complex problems, enabling me to approach them with greater insight and clarity. Research is never smooth sailing, and for me, the most challenging period of my doctoral studies coincided with the global COVID-19 pandemic. With his support and encouragement, I survived the tough period and made essential progress afterward. Without his mentorship, I could not have finished this dissertation and proceeded to the next stage of my career.

I would like to thank Negin Rahimi for her assistance during the early stages of my Ph.D. program. Collaborating with her was invaluable in building my research vision and attitude. I am deeply grateful to committee members Mohit Iyyer and Jeffrey Dalton, whose valuable comments and suggestions greatly enhanced the clarity and consistency of my dissertation. Special thanks to Jeff for his meticulous edits and the insightful questions he posed during my defense. Also, I thank Hamed Zamani for his feedback on my research project, which finally became part of the dissertation.

I would like to thank the staff in our lab and the college. I am particularly grateful to Dan Parker for the technical support. My appreciation also goes to Kate Moruzzi, Jean Joyce, and Michael Schwendenmann for their dedicated administrative and logistical support. I thank our Associate Director of Graduate Programs, Eileen Hamel, Graduate Programs Assistant, Kyle Skemer, and Senior Academic Advisor, Elizabeth Parolski, for their support at every milestone throughout my entire program.

I extend my thanks to current and former labmates, listed here in alphabetical order: Ali, Alireza, Chen, Chris, Hansi, Julian, Keping, Mahta, Nazanin, Qingyao, Rab, Shahrzad, Sheikh, Tanya, Yaxin, Yen-Chieh, and Youngwoo. Our discussions have provided numerous insights, and sharing our life experiences has been truly enjoyable. I especially thank Puxuan for being a collaborator, a roommate, and a friend throughout my Ph.D. study.

At last, I would like to express my deepest gratitude to my family for their immense love and unwavering support. My heartfelt thanks go to my grandparents, who passed away during my Ph.D. study. Their wisdom, love, and encouragement have always been guiding lights in my life. I am also profoundly grateful to my parents, who have stood by me through both good and bad days. Their constant encouragement and unwavering presence have been my greatest blessings, and I am incredibly fortunate to have them in my life.

This work was supported in part by the Center for Intelligent Information Retrieval, and in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract #2019-19051600007 under University of Southern California (USC) subcontract #124338456, in part by the ODNI IARPA via Air Force Research Laboratory (AFRL) contract #FA8650-17-C-9116 under USC subcontract #94671240, and in part by the AFRL and IARPA via contract #FA8650-17-C-9118 under Raytheon BBN Technologies Corporation subcontract #14775. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

ABSTRACT

MODELING CROSS-LINGUAL KNOWLEDGE IN MULTILINGUAL INFORMATION RETRIEVAL SYSTEMS

SEPTEMBER 2024

ZHIQI HUANG

B.Sc., SUN YAT-SEN UNIVERSITY

M.A., UNIVERSITY OF MARYLAND COLLEGE PARK

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor James Allan

In many search scenarios, language can become a barrier to comprehensively fulfilling users' information needs. An Information Retrieval (IR) system equipped with an extra component of language translation is capable of mapping words in different languages, enabling it to retrieve documents according to the user's query regardless of the language in which the query and documents are expressed. Effectively incorporating multilingual knowledge is the key to building the translation component. Such knowledge can be obtained from dictionaries, machine translation modules, or multilingual pre-trained language models. For these different forms of multilingual knowledge, we present cross-lingual knowledge injection, transfer, and language debiasing techniques to enhance the effectiveness of Cross-lingual Information Retrieval

(CLIR) and Multilingual Information Retrieval (MLIR). Specifically, by utilizing multilingual knowledge at various levels—from individual word translations to parallel and non-parallel corpora—we develop new model architectures and training goals tailored for information retrieval tasks across diverse linguistic settings.

First, we introduce a mixed attention Transformer layer, which augments mutually translated words between query and document into the attention matrix and investigates its effectiveness on CLIR tasks. Next, we study cross-lingual transfer in the IR models and demonstrate a knowledge distillation framework to address the data scarcity problem in model training and improve retrieval effectiveness involving low-resource languages. Then, we focus on a special setting in MLIR, where the query is in one language, and the collection is a mixture of languages. To address the problem of inconsistent ranking results between languages, we design an encoder-decoder model that maps document representations from different languages into the same embedding space. We also present a decomposable soft prompt to capture unique and shared properties across languages.

Finally, we introduce a language debiasing method to identify and remove linguistic features from a multilingual embedding space. This approach significantly diminishes the necessity for parallel data in constructing MLIR models, allowing for using non-parallel data instead. By reducing language-specific factors from the training process, we improve the retrieval effectiveness for all linguistic settings in retrieval tasks (e.g., monolingual, cross-lingual, and multilingual), thereby facilitating language-agnostic information retrieval.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

As the primary medium of communication, language plays a pivotal role in designing and implementing Information Retrieval (IR) systems. Retrieval under a monolingual setting assumes the query and documents are in the same language. When documents are desired in another language, it is often reasonable to expect the user to be able to formulate a query in that language. Nevertheless, there are important needs that cannot be satisfied by monolingual retrieval systems (Oard and Diekema, 1998):

- In multilingual societies, it is important to access information in different languages, especially for tasks like government services, healthcare, and education.
- As English is the de facto standard language of science and technology, it also creates a language barrier for millions worldwide who are not fluent in English. It is essential to bridge this language gap and make information more accessible to these people, especially for low-resource language speakers.
- The internet is a multilingual space, with content available in many languages. Accessing information regardless of language promotes knowledge sharing and cultural exchange.

Breaking the limitation of language, Cross-lingual Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR) can provide a more comprehensive fulfillment of the user's information needs. CLIR is the process of retrieving information written in a language different from the user's query. In a particular CLIR task,

a language pair is given where the user submits the query in one language and the systems respond by retrieving documents in another language. Compared to CLIR, MLIR has a more general linguistic setting. It is concerned with retrieval from a collection where documents in multiple languages co-exist and need to be retrieved in response to a query. Generally, these tasks require matching queries and documents in different languages. In addition to the ranking component, the retrieval models for CLIR or MLIR need to possess some knowledge of translation to map the vocabulary of the query language to that of the documents' language. Therefore, retrieval effectiveness depends on knowledge of query document matching and the ability to bridge the translation gap between query and document. The translation knowledge can be acquired from different resources, such as a dictionary, a machine translation module, or multilingual word embeddings.

Based on the Transformer architecture (Vaswani et al., 2017), pre-trained language models (PLMs) such as BERT (Devlin et al., 2019) are capable of encoding linguistic and factual knowledge into their deep neural network parameters. Fine-tuned on task-specific data, PLMs offer great success for many downstream tasks, including document ranking. The multilingual versions of PLMs (mPLMs), such as mBERT and XLM-R (Conneau et al., 2020), provide the possibility of jointly learning representations for multiple languages with the same model. Because tokens in different languages are projected into the same space, these models can also be adopted as the source of translation knowledge. And like the monolingual setting, fine-tuning mPLMs with multilingual retrieval data enables information retrieval across languages. However, the models in a multilingual setting are not achieving the same level of performance as those in a monolingual setting. Even with the help of powerful mPLMs, challenges like the persistence of translation gaps, data scarcity on low-resource languages, and ranking inconsistency across languages still exist for building effective CLIR or MLIR models.

Under the monolingual retrieval setting, because the query and document use the same lexical inputs, it is easier for a model to identify words that co-occurred in both query and document. The co-occurrence becomes mutually translated words when the query and document are in different languages. We observed that mPLMs tend to map query terms into the target language's related terms – i.e., terms that appear in a similar context – in addition to or sometimes rather than its synonym translations. Studies on CLIR (Bonab et al., 2020; Nie, 2022) have shown that the translation gap plays a significant role in the suboptimal results of neural CLIR models. Therefore, re-introducing the external translation knowledge into the neural CLIR models effectively reduces the translation gap.

Further, unlike the English-to-English retrieval task, where many resources are available for model training, the scarcity of retrieval data in other languages, especially in low-resource languages, makes it challenging to build multilingual retrieval models. Transferring retrieval knowledge learned from English retrieval data to other languages is a promising solution to relieve the data scarcity problem.

Moreover, in multilingual search situations, the retrieval collection consists of documents in multiple languages. The distribution of relevant documents for a specific query often varies across languages. For example, consider a situation where a user is searching for academic articles on "climate change adaptation strategies" from a multilingual collection. Due to regional research focuses, compared to English articles, the Spanish documents might contain more detailed case studies from South America, while the French documents could be rich in strategies used in Francophone Africa. It is important for the retrieval model to evaluate and rank documents irrespective of their written language, even when the query language is also one of the languages in the collection. This ensures that users receive the most relevant results, regardless of linguistic barriers, and promotes an inclusive and comprehensive search experience for a diverse user base.

To address the challenges mentioned above, we design new model architectures and training objectives to effectively incorporate different forms of multilingual knowledge, from which we will show results in the improvement of both CLIR and MLIR system performance.

Note that there are some studies (Sun and Duh, 2020; Yang et al., 2020; Zhang et al., 2021, 2023) focusing on monolingual retrieval tasks on multiple languages, which also use the term "multilingual retrieval" in their task definition. To avoid term confusion in this dissertation, we refer to these studies as monolingual retrieval or *in-language* retrieval. In our definition, "multilingual retrieval" refers to a retrieval task involving multiple languages in either query or document, or both.

1.1 Translation Knowledge Injection

From monolingual to cross-lingual, the exact matching of terms in documents with those in queries becomes translations in two languages. When using mPLMs for document ranking, multilingual models tend to map query terms into the target language's related terms—i.e., terms that appear in a similar context—in addition to or sometimes rather than translations (Zhan et al., 2020). The translation misalignment weakens the signal of "exact match" in the cross-lingual context, creating the translation gap between the query and document in the retrieval task.

In Chapter 3, we study the translation gap in cross-lingual document ranking models. When fine-tuning with cross-lingual relevance data, we inject word-level translation knowledge as a fixed attention mechanism into the CLIR model. More specifically, we leverage the external knowledge in the form of a translation table: a look-up table that provides translation probabilities for a pair of words in two languages. Our novel network module uses the translation table to create an attention matrix and parallels it with the Transformer's multi-head self-attention – both in the training and inference phase – to improve the model's cross-lingual understanding.

We refer to our extended component as Mixed Attention Transformers (MAT) and create MART (MAT+BERT), a sandwich-like architecture to embed MAT into the multilingual BERT (mBERT) model. Encoding the translation knowledge into an attention matrix enables the overall architecture to focus on the mutually translated words in the input sequence.

Our analysis of the representation shows that applying MAT layers successfully reduces the translation gap by increasing the cosine similarity of representation of mutually translated terms in query and document. We also explore the effectiveness of various external knowledge sources and show the significant gain we get from MART on the CLIR task. MAT is a generalized architecture capable of capturing any form of lexical mapping, and it can be integrated with any transformer-based architecture.

- *Contribution 1.1.* We present a novel Mixed Attention Transformer (MAT) network to leverage external translation knowledge for CLIR tasks. As a layer component, we further design a hybrid architecture to embed MAT into the Transformer model.

- *Contribution 1.2.* We perform extensive experiments on ten different language pairs for CLIR training and evaluation, three different resources to obtain translation knowledge, and different qualities of translations based on available translation resources for language pairs. Our experimental results demonstrate the effectiveness of external knowledge sources and the significant improvement of the MAT-embedded neural re-ranking model over strong baselines on the CLIR task. In terms of mean Average Precision (mAP), our proposed model outperforms the neural baseline by 8% on high-resource languages and 12% on low-resource languages.

1.2 Cross-lingual Transfer via Knowledge Distillation

Compared to English-to-English retrieval, CLIR and MLIR tasks usually suffer from the data scarcity problem where insufficient queries with reliable relevance judgments are available for model training (Sasaki et al., 2018). This issue becomes more severe when the retrieval tasks involve low-resource languages. Instead of directly training models on target language using retrieval data, we study how to extract ranking knowledge from a well-trained English retrieval model and transfer it into the target languages.

In Chapter 4, we introduce a cross-lingual transfer framework based on knowledge distillation to transfer English retrieval models to another language. More specifically, we first train a bi-encoder English retrieval model (Khattab and Zaharia, 2020) as the teacher, using the multilingual pre-trained encoder. It learns how to do query-document matching from abundant English retrieval data, such as MS MARCO passage ranking dataset (Nguyen et al., 2016). Suppose the task is to search English documents with non-English queries. We then reuse the teacher's document encoder and train a new student query encoder. The student model distills retrieval knowledge from the teacher model through a task of cross-lingual token alignment. Unlike previous approaches (Gritta and Iacobacci, 2021; Li et al., 2022), which align tokens using a rule-based algorithm, we conceptualize training as an optimal transport problem where the cost matrix is the token-level cosine distance, and the optimal transportation plan acts as a soft token alignment. By separating the learning of cross-lingual knowledge from retrieval knowledge, the cross-lingual transfer process only needs *bitext data* for training. Bitext [1], also known as parallel text or parallel corpora, contains translations of the same or comparable document in two or more languages, aligned at least at the sentence level. Unlike retrieval data, which typ-

[1] `https://en.wikipedia.org/wiki/Parallel_text`

ically requires human judgment, bitext data can be extracted through automated algorithms (El-Kishky et al., 2020a; Heffernan et al., 2022). Therefore, our approach greatly expands the languages capable of implementing CLIR tasks and covers some low-resource languages.

- *Contribution 2.1.* Focusing on low-resource languages, we present OPTICAL: Optimal Transport distillation for Cross-lingual information retrieval.

- *Contribution 2.2.* We show that in terms of mAP, our proposed method significantly outperforms several strong baseline methods on four low-resource languages from different language families, including a 13.7% improvement over a method based on neural machine translation. Further analysis demonstrates that the knowledge distillation step in OPTICAL is an effective and data-efficient method to transfer retrieval knowledge from monolingual into cross-lingual settings.

- *Contribution 2.3.* We extend our language transfer technique to monolingual retrieval tasks other than English. The experimental results show that our method effectively improves retrieval performance in a group of 16 languages. When used as a module in an ensemble retrieval system, it helped to achieve one of the top submissions in the MIRACL [2] (Multilingual Information Retrieval[3] Across a Continuum of Languages) leaderboard.

1.3 Language Prompt for Multilingual Knowledge Transfer

An MLIR system can retrieve documents based on explicit queries formulated by a human using natural language, regardless of the language in which the documents and

[2] `https://www.wsdm-conference.org/2023/program/wsdm-cup`

[3] Note that this challenge is made of monolingual retrieval tasks in 16 languages which are different from the concept of MLIR used in this dissertation.

the query are expressed. Even though CLIR and MLIR are tightly coupled, effective MLIR models should overcome additional major challenges. For instance, instead of one pair of languages between query and document, the translation component in the MLIR model needs cross-lingual knowledge for multiple language pairs. Meanwhile, it requires the system to perform fairly across languages, that is, to eliminate the language factors when ranking documents against a query.

In Chapter 5, we study a special case of MLIR, where the query is always written in English, and the collection is a mixture of languages, known as the one-to-many setting. Following the idea of cross-lingual transfer, we develop KD-SPD, a multilingual dense retrieval model based on knowledge distillation (KD) and soft prompt decoder (SPD) for the MLIR task. Using an encoder-decoder architecture, our model implicitly "translates" the representation of documents in different languages into the same language embedding space as the query. And the decoding is also a knowledge distillation process guided by a teacher model built for monolingual retrieval in English. Again, the distillation training is supported by bitext data instead of retrieval data between each language pair of query documents. We hypothesize that although different languages possess unique properties such as distinct grammar or vocabulary, they also have common traits for expressing similar meanings. To capture both unique and shared features, the decoder input of KD-SPD uses a decomposable soft prompt (Wang et al., 2023) derived as the product of a low-rank language-specific matrix and a matrix for the shared feature. Through joint training across multiple languages, we observe that the learned prompts are capable of reducing language bias and possessing the transferable capacity to generalize to unseen languages.

- *Contribution 3.1.* We investigate the one-to-many setting in the MLIR task and develop a new model combining knowledge distillation and soft prompt decoding to project the representation of documents in different languages into the same language embedding space.

- *Contribution 3.2.* We conduct extensive experiments on MLIR datasets with a total of 15 languages from diverse linguistic families and discover that, in terms of mAP, our proposed method significantly outperforms several strong baselines, including 20.2% improvement over multilingual dense passage retriever (mDPR) (Zhang et al., 2021) and a 9.6% improvement over a multilingual knowledge distillation method from Sentence-BERT (Reimers and Gurevych, 2020).

1.4 Language Concept Erasure in Dense Retrieval Models

From our previous studies in CLIR and MLIR, we find that retrieval knowledge is distinct from linguistic knowledge and can be transferred across different languages. Utilizing parallel corpora, we facilitate such transfer through knowledge distillation frameworks between multiple languages.

The separation of different types of knowledge during retrieval modeling inspires us to explore a universal search engine that can effectively retrieve relevant information across all linguistic contexts. In the paradigm of natural language understanding (NLU), this is similar to embedding disentanglement (Tiyajamorn et al., 2021; Wu et al., 2022), where sentence embeddings are viewed as the combination of semantic (meaning) embedding and language-specific (syntax or idioms) embedding. For a particular task, it is preferable to concentrate model training on the semantic embedding while deliberately excluding the language-specific part. This strategy aims to enhance the model's ability to function effectively across diverse linguistic environments, improving its universal applicability and performance.

Therefore, to build a language-agnostic dense retrieval model, in Chapter 6, we introduce a multi-task learning framework to reduce linguistic influence within the representation space. Given multilingual inputs, we consider language as a predictable concept tied to each input utterance and leverage a conceptual erasing task to ob-

scure the language labels within the output representations. More specifically, during training, we calculate the cross-correlation matrix between the vectors produced by a dense retriever and the language labels for each training batch. By minimizing the mean correlation values across batches, this task prevents all linear classifiers from detecting the language label, thus reducing linguistic features from the representations.

While the primary task is learning retrieval knowledge, language concept erasure serves as an auxiliary task to drive the model toward generating language-agnostic representations. Concurrently, the retrieval task helps to prevent trivial solutions in the concept erasure task by ensuring that the model maintains a meaningful representation throughout the training process. Since the language label or linguistic feature is an inherent attribute existing in any context of a certain language, the proposed method is capable of operating on multilingual non-parallel corpora, effectively diminishing the necessity for parallel data for constructing a language-agnostic retriever. The dense retrieval models developed using our framework exhibit reduced language bias in MLIR tasks, especially between English, which carries the retrieval knowledge during training, and other languages. Benefiting from language-agnostic representations, these models also demonstrate a substantial improvement in monolingual and cross-lingual retrieval tasks.

- *Contribution 4.1.* We develop **Lan**guage **C**oncept **E**rasure for Language-Agnostic Dense **R**etrieval (LANCER). We formulate the goal of removing language-specific signals from the model's representation as preventing any linear classifier from detecting the language label of the model inputs. In conjunction with retrieval training through ranking loss utilizing English datasets, we devise a multi-task learning framework that drives the model toward language-agnostic representations using multilingual non-parallel context.

- *Contribution 4.2.* Applied to dense retrieval model training based on the different multilingual pre-trained encoders, in terms of nDCG@10, LANCER significantly improved retrieval performance on MLIR with both query and document being multilingual (known as many-to-many setting), including 36.6% improvement over mDPR and 27.6% improvement over mContriever (Izacard et al., 2021). Evaluated on the MIRACL dataset (in-language retrieval tasks), on average, it improves mDPR and mContriever by 13.6% and 32.5%, respectively.

1.5 Summary

In this dissertation, we investigate different techniques to model cross-lingual knowledge in CLIR and MLIR systems. First, in Chapter 3, we obtain word-level translations from parallel corpora and reintegrate them into neural document rankers. Then, in Chapter 4, we present a knowledge distillation method to incorporate cross-lingual knowledge directly from parallel sentences. Meanwhile, we introduce a framework for building models that leverage well-established English dense retrieval models and transfer retrieval knowledge from English to the target language in a cross-lingual setting. Extending from CLIR to MLIR, in Chapter 5, we investigate the task of searching a multilingual collection using English queries and mapping the representation of documents from different languages into the same English retrieval space to reduce language bias in multilingual document ranking. Lastly, different from using English as the pivot language and transferring multilingual to monolingual embedding space (the approach of KD-SPD), Chapter 6 leverage the concept erasure method to reduce the language-specific features for all language contexts in the embedding space, achieving language-agnostic retrieval.

Our approaches are inspired by the extensive research conducted in the fields of CLIR and MLIR. In the following chapter, we outline the background and related works to this dissertation.

CHAPTER 2

RELATED WORK

We start in Chapter 2 by discussing the different forms of cross-lingual knowledge employed in CLIR studies. Then, we present an overview of neural ranking models and their adaptations to multilingual retrieval tasks. Lastly, we discuss the challenge from CLIR to MLIR and the previous works related to MLIR.

2.1 Cross-lingual Knowledge

From a modeling perspective, when a query and candidate documents are in different languages, a translation component is required in addition to the ranking component to overcome the vocabulary mismatch between the query language and the document language. In the CLIR task, the translation occurs between two specific languages, whereas in the MLIR task, translations are needed for multiple pairs of languages. This translation process can be executed either explicitly, using dictionaries or machine translation modules, or implicitly, through techniques such as bilingual word embeddings and multilingual pre-trained language models. Although these approaches vary, they all strive to enable the retrieval system to effectively integrate cross-lingual knowledge. In the following sections, we will discuss the details of these translation components.

2.1.1 Explicit Translation

Dictionary-based Translation. As a long-standing problem in IR, early CLIR approaches directly adopt the cross-lingual knowledge from a dictionary. The cross-lingual setting is first explored in SMART (System for the Mechanical Analysis and

Retrieval of Text). Salton (1970) discussed the possibility of searching English documents against German queries and vice versa. To tackle the vocabulary mismatch, queries and documents in different languages are converted into predefined concepts using a synonym dictionary, known as a bilingual thesaurus. These representations are then reduced to "concept vector" forms, which allow for effective comparison between different languages. Later, the appearance of bilingual machine-readable dictionaries (MRDs) provides a groundwork for constructing query (or document) translations in CLIR tasks. Early systems translate queries on a word-by-word basis: for each query term, select the first translation offered by the dictionary. This strategy exploits the fact that the most commonly used translation is listed first in some bilingual dictionaries (Davis and Dunning, 1995; McDonnell et al., 1995). Then Ballesteros and Croft (1997) point out the limitation of word-by-word translation and present phrasal translation and query expansion methods that can greatly reduce the error associated with dictionary-based translation.

Moreover, in most MRDs, a word in one language can map to multiple words (polysemous) in another language. Some senses from MRDs translation are inappropriate to the query and introduce ambiguity, eventually leading to poor retrieval effectiveness (Ballesteros and Croft, 1998). Techniques exploiting term co-occurrence statistics can disambiguate the translations from dictionaries (Adriani, 2000; Gao et al., 2002; Liu et al., 2005). In this context, the underlying assumption for utilizing term co-occurrence data is that accurate translations of individual query terms will typically appear together as part of a sublanguage (Grishman et al., 1986), whereas incorrect translations will not. Essentially, this method aims to identify the most probable translation for a specific query by analyzing the term co-occurrence pattern within a representative text collection, e.g., the World Wide Web (Maeda et al., 2000) or a monolingual corpora (Ballesteros and Croft, 1998; Gao and Nie, 2006).

Corpus-based Approach. Parallel and comparable corpora are another source of learning cross-lingual knowledge. The knowledge extracted from these corpora is commonly used in cross-language information retrieval to translate queries. The basic technique involves token alignment of bilingual text corpora, producing a set of transition probabilities for each term in a given query. A structural collection of these probabilities is a *translation table*. Known as Statistical Machine Translation (SMT), the translation table is learned based on statistical models. The early approaches for extracting translations from the parallel corpus are based on the EM algorithm and the Hidden Markov Model (HMM). The most popular tool for extracting word or phrase alignment from the parallel corpus is GIZA++(Och and Ney, 2003). As an ensemble of IBM alignment models (Brown et al., 1993), the training pipeline of GIZA++ relies on multiple iterations of IBM Model 1, Model 3, Model 4, and the HMM alignment model (Vogel et al., 1996).

Another category of approach is context vector projection. Relying on an existing dictionary, these approaches extract more translations from the comparable corpus. First, the vector for each word is built based on all words co-occurring in a context window. According to the dictionary, the source-language word vector is then projected to the target-language word vector. The word vector in the target language is ranked based on the similarity to the projected vector to identify translations (Déjean et al., 2002; Sadat et al., 2003). Besides co-occurrence in a context window, Gaussier et al. (2004) employed latent semantic analysis for deriving word vectors by presenting a geometric view of translation extraction. Dependency trees are also used for modeling context vectors of words (Garera et al., 2009). Hazem and Morin (2014) combined window- and syntax-based contexts of words to build word vectors. Deriving correlations between source–target word pairs, Rahimi et al. (2016) proposed a language modeling approach to extract translations from comparable corpora without relying on a dictionary.

Machine Translation. Another way to model cross-lingual knowledge for retrieval tasks is to use an off-the-shelf Machine Translation (MT) module independent from the retrieval module. In the corpus-based approach, we already discussed Statistical Machine Translation (SMT) using statistical models whose parameters are derived from the analysis of bilingual text corpora. Herein, we review another type of MT model, Neural Machine Translation (NMT), and its performance in CLIR systems. NMT is generally an end-to-end machine translation model that builds on a neural network. Since integrating the attention mechanism, NMT systems have seen a remarkable improvement in translation quality. Most commonly, an attentional NMT model consists of three components: (a) an encoder that computes a representation for each source sequence; (b) a decoder that generates one target symbol at a time; (c) the attention mechanism that computes a weighted global context concerning the source and all the generated target symbols (Klein et al., 2017). Named the sequence-to-sequence (Seq2Seq) model, the network architecture of an NMT model is usually a stack of Recurrent Neural Networks (RNNs) or Transformer layers (Bahdanau et al., 2014; Vaswani et al., 2017). Studies have shown that a well-performed NMT model depends on extensive language resources for training (Ott et al., 2018). It learns cross-lingual knowledge from the corpus and integrates it into the network parameters. Yao et al. (2020b) employed the Transformer-based NMT model as the query translation module and showed that with top translation output, NMT outperforms SMT in terms of translation quality and leads to better retrieval performance. However, Bonab et al. (2020) showed that with limited data available for training, NMT struggles to match the performance of SMT. Concerns have been pointed out for applying NMT as a module for query translation (Sarwar et al., 2019; Yao et al., 2020a):

- As a component in CLIR, query translation has to be real-time, while the advanced NMT model built with multiple network layers may fail to cope with the requirement of processing speed.

- NMT models generally pay great attention to syntactic structure, which is less important when translating queries for retrieval. The focus on the fluency of the output may reduce the accuracy and coverage of the translation.

- Compared to SMT, NMT is less efficient in learning cross-lingual knowledge. Its performance is highly dependent on the vast amount of training data, making SMT a preferable choice for languages with limited resources.

2.1.2 Implicit Translation

Utilizing Word Embeddings. Incorporating cross-lingual knowledge through dictionary-based translation, corpus-based translation, and machine translation modules typically entails executing explicit translations. To search the documents, CLIR or MLIR systems employ a two-stage pipeline: translate first, then retrieve. In contrast to explicit translation, some research efforts focus on performing translation implicitly. Bilingual word embeddings create the opportunity to skip the translation step. As discussed in the previous subsection, early studies in using dictionaries as a translation medium explored the use of predefined "concept vectors" to represent words, converting the query and document into vector forms (Salton, 1970, 1973). This is also knowns as the one-hot encoding of word embeddings. Different from predefined vectors, the learned word embedding techniques, on the other hand, depend on a training task to derive a low-dimensional vector (compared to the vocabulary size) for each vocabulary term (Mikolov et al., 2013a; Pennington et al., 2014). These vectors capture semantic and syntactic similarities between the corresponding words, reflecting their relationship within the embedding space (Mikolov et al., 2013b). Similar to single-language word embedding, bilingual word embeddings represent lexicons

of different languages in a shared embedding space. As a result, query-document matching can be performed in a shared vector space for two languages, where words that have similar meanings in two different languages are mapped close to each other. The underlying assumption is that the embedding space contains cross-lingual knowledge. Utilizing bilingual word embeddings to represent queries and documents serves as an implicit translation approach. One of the earliest works in this direction is from Vulić and Moens (2015), who proposed a model to learn bilingual word embeddings using document-aligned comparable data. Once all the words in both languages are represented in a shared space, they computed query and document representations using the distributional semantics model to calculate their matching score based on the cosine similarity metric. Bonab et al. (2020) assessed the effectiveness of several bilingual word embeddings under a cosine similarity-based scoring framework for retrieval and found bilingual word embedding can bring similar pairs of words in two languages close together but often keeps the words that are translations of each other farther apart than expected. This is because cross-lingual word embeddings are learned from surrounding words of a target word but not from the translation of that word. They referred to this phenomenon as the *translation gap* and proposed a smart shuffling approach to include translation knowledge into word embeddings, improving CLIR performance.

Multilingual Pre-trained Language Models. One of the main limitations of static word embeddings or word vector space models is that words with multiple meanings are conflated into a single representation. In contrast, contextual language models like ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), and RoBERTa (Liu et al., 2019) employ deep neural networks (multiple layers of LSTMs or Transformers) and attention mechanisms to learn context-dependent representations of input tokens based on their surrounding context. Instead of optimizing for a particular task, these models are focused on learning contextualized representation

through pre-training tasks. BERT and RoBERTa are pre-trained on the Masked Language Modeling (MLM) objective, also known as the Cloze task (Taylor, 1953). Raffel et al. (2020) introducing a unified encoder-decoder framework, Text-to-Text Transfer Transformer (T5), that converts all text-based language problems into text-to-text pre-training tasks. These pre-training tasks effectively capture general linguistic patterns and greatly enhance the performance of various natural language processing (NLP) tasks. Extending the pre-training task to the multilingual setting creates the multilingual versions of pre-trained language models (mPLMs), such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021). Unlike BERT, which is trained on English Wikipedia and the Toronto Books Corpus, mBERT is trained on up to 104 languages from Wikipedia. Besides datasets, additional cross-lingual pre-training objectives are also applied to improve the pre-training of multilingual language models. Based on the RoBERTa model, XLM-R includes a translation language modeling (TLM) objective to predict a masked English word within a pair of parallel sentences. The development of mPLMs allows for jointly learning contextualized representations for multiple languages within the same vector space. Studies have shown that mPLMs possess multilingual knowledge and are capable of performing cross-lingual transfer in their representation space (Pires et al., 2019; Wu and Dredze, 2019). By fine-tuning using specialized task data, the mPLMs can be applied to various multilingual tasks, including CLIR and MLIR, demonstrating substantial improvements over static word embedding approaches.

Our approaches for incorporating cross-lingual knowledge utilize both explicit and implicit translation methods. In this dissertation, mPLMs serve as the basic textual encoder for all the methods discussed. Despite the high quality of the contextualized representations, we focus on three limitations of applying mPLMs to CLIR and MLIR tasks: (1) Similar to bilingual word embeddings, the translation gap still exits in mPLMs (Zhan et al., 2020); (2) Fine-tuning mPLMs requires cross-lingual retrieval

data, which are often costly to acquire, especially for low-resource languages (Sasaki et al., 2018); (3) mPLMs exhibit performance variations across different languages in many downstream tasks, including retrieval. This phenomenon, known as language bias, hinders the ability of mPLM-based retrieval systems to rank documents fairly across languages (Wu and Dredze, 2020).

To bridge the translation gap, we combine multilingual contextualized representation with dictionary knowledge (Chapter 3). Then, instead of using retrieval data for training, we introduce a knowledge distillation framework for cross-lingual transfer via bitext data (Chapter 4). Finally, we present two methods to minimize language bias in mPLMs-based retrieval models: (i) utilize English as a pivot language and form language features as a soft prompt to map representations from various languages into the English retrieval space. (ii) erase language-specific features from model output and focus the retrieval training on the language-agnostic representations. (Chapter 5 and Chapter 6).

2.2 Neural Ranking Models

2.2.1 Model Architecture

The development of the neural ranking model ties in with the approaches of representing the query and document in vector space. In earlier works, based on deep learning ideas, words in both the query and document are represented using static word embeddings (Mikolov et al., 2013a). To contextualize a bag of words, kernel methods are employed to construct query-document interactions. Guo et al. (2016) proposed the DRMM, which uses different matching histogram strategies to convert the pair-wise similarity matrix between query terms and document terms. Starting from the same similarity matrix, K-NRM (Xiong et al., 2017) applied kernel pooling to transform word-word interactions into ranking features. Instead of static word embeddings, Li and Cheng (2018) took an adversarial learning approach to jointly

learn language alignment through translation knowledge and cross-lingual matching using relevance judgments. Bonab et al. (2020) proposed translation-oriented bilingual word embeddings and combined them with DRMM matching model for CLIR tasks.

To take advantage of the contextualized representation, one paradigm to incorporate PLMs for document ranking tasks follows a cross-encoder architecture. A special `[CLS]` token is prepended to the input sequence to support the downstream application. Because embedding of the special token is contextualized based on other tokens in the input sequence, once fine-tuned, they are effective across various tasks, including retrieval tasks. For the cross-encoder architecture, the model takes the query document concatenation as the input. An embedding produced from the `[CLS]` token is fed into a feed-forward perceptron layer to produce a ranking score (Dai and Callan, 2019; MacAvaney et al., 2019; Nogueira and Cho, 2019). However, retrieval models that employ cross-encoders are computationally intensive and typically depend on a lexical-based sparse retrieval method as an initial step to identify relevant information. Thus, they are often known as the reranking model or re-ranker. The dense retrieval model based on bi-encoder architecture is introduced to overcome the sparse retrieval bottleneck. In this context, the model usually contains two encoders that process the query and document separately. Khattab and Zaharia (2020) proposed a late interaction dense retriever named ColBERT, which delays the interaction until the scoring function. The scoring function applies the *maxsim* operation on each query token to softly search against all document tokens to find the best token that reflects its context and then sums over all the query tokens. Karpukhin et al. (2020) proposed Dense Passage Retriever (DPR) that further simplifies the representations of query and document into two vectors, respectively, and employs the dot-product as the scoring function. Since the query and document are encoded separately, the document collection can be indexed offline, and a search of the top-K documents

is equivalent to finding the K nearest document embeddings to the query embedding (Johnson et al., 2019). Dense retrieval based on bi-encoders can be applied autonomously to search the entire collection or integrated with reranking models to generate preliminary ranking lists. This approach has already demonstrated efficacy in English retrieval tasks (Gao and Callan, 2022; Xiong et al., 2021).

Both cross-encoder and bi-encoder models for English can be configured for CLIR and MLIR by substituting the underlying PLMs with their multilingual counterparts. Yu et al. (2021) explored the performance of cross-encoder on CLIR tasks by combining the masked language model pre-training task with the retrieval fine-tuning task. Nair et al. (2022) and Lawrie et al. (2023) extended ColBERT to the multilingual setting by replacing BERT with XLM-R and evaluating on CLIR and MLIR datasets. Utilizing mPLMs is an effective method for incorporating cross-lingual knowledge when constructing CLIR or MLIR models. Nevertheless, this strategy also inherits the limitations of mPLMs, as mentioned in the prior subsection.

2.2.2 Model Training

In general, the training of neural retrieval models requires queries and documents labeled with relevance scores. This can either be explicit relevance judgments (e.g., human-annotated datasets) or implicit feedback (e.g., click data). The most common loss function used in training cross-encoders is the binary cross-entropy loss for binary relevance tasks or a ranking loss such as pairwise hinge loss if the model needs to learn from a ranking context. The model aims to minimize the loss to correctly predict the relevance of the document to the query (Nogueira and Cho, 2019). Similar to cross-encoders, dense retrieval models are trained on datasets of query-document pairs. However, the training sets must be carefully designed to include negative examples (non-relevant documents) alongside positive examples to effectively train the dense representations. Negative examples are commonly sampled from the ranked

list of lexical-based retrieval methods, such as BM25 (Zhan et al., 2021). In-batch data can also be used to expand on negative examples for each training instance. Xiong et al. (2021) proposed an asynchronous learning mechanism that selects hard training negatives globally from the entire corpus, using the previous index generated by the previous model checkpoint. After retrieval data is collected, contrastive loss, such as margin-based loss, is commonly employed. These loss functions typically involve a positive pair (query and relevant document) and one or more negative pairs (query and non-relevant documents), aiming to maximize the distance between the query's embedding and the embeddings of non-relevant documents while minimizing the distance to relevant document embeddings.

A straightforward approach to building neural retrieval models for CLIR and MLIR tasks follows the same training paradigm of retrieval tasks in English, substituting the English retrieval data with the retrieval data in target languages. However, in practice, the distribution of resources for pre-training mPLMs is not uniform across all languages, leading to an imbalance in the data available for different languages. This disparity has been observed to result in a performance gap between high-resource and low-resource languages across various downstream tasks. Consequently, commonly known as the language bias issue, mPLMs tend to exhibit enhanced effectiveness in languages with abundant resources, while their performance in languages with limited data remains suboptimal (Wang et al., 2020; Wu and Dredze, 2020). Retrieval models built on such mPLMs models can inherit the language bias. Moreover, because English is the dominant language on the internet, digital devices, and academia (Blodgett et al., 2020), compared to the scale of labeled query-document in English, the training of CLIR and MLIR models also faces the data scarcity problem (El-Kishky et al., 2020b). The limited availability of relevance judgments within target languages severely restricts models' ability to develop effective retrieval knowledge (Litschko et al., 2022a,b). Prior studies focused on building CLIR and MLIR

datasets for better training or evaluation. Datasets based on human annotations, such as NeuCLIR (Lawrie et al., 2024) and LAReQA (Roy et al., 2020), have been proposed for model evaluation. To mitigate the data scarcity in model training, large-scale synthetic training data generation involves two main simulation strategies: translating English retrieval datasets into target languages or building pseudo-labels using a corpus in target languages. For example, Sasaki et al. (2018) proposed a large cross-lingual retrieval collection, WikiCLIR. It uses the title of articles in target languages linked from Wikipedia pages as the query to simulate relevance. Bonifacio et al. (2021) built a multilingual passage ranking dataset, mMARCO, by translating the queries and passages in MS MARCO into the target language using Neural Machine Translation (NMT) models. Thakur et al. (2023) leveraging the recent advances in generative and autoregressive Large Language Models (LLMs) to generate queries in target languages with minimal supervision.

This dissertation focuses on neural retrieval models, covering both cross-encoder document reranker and bi-encoder dense retriever. Chapter 3 focuses on cross-encoder document reranker for the CLIR task. With limited cross-lingual retrieval labels for training, we introduce word-level translation knowledge into mPLMs to address the translation gap. For bi-encoders, instead of simulating retrieval labels in target languages to support the model training, we suggest employing cross-lingual knowledge transfer using bitext data to tackle the data scarcity issue. Chapter 4 studies knowledge transfer via cross-lingual token alignment for multi-vector dense retrieval models, and Chapter 5 extends such knowledge transfer to multiple language pairs. Based on the findings of knowledge transfer in CLIR and MLIR tasks, Chapter 6 explores language concept erasure, an approach to removing language-specific signals in multilingual embedding space, and encouraging the training of dense retrieval models to focus on language-independent retrieval knowledge.

2.3 From Cross-lingual to Multilingual

While CLIR involves searching between two distinct languages, MLIR, by definition, allows for any language to be used as input for both queries and documents. Given that there are more than 7000 languages worldwide [1], in practice, the language setting of MLIR is narrowed down to a set of query languages and a set of document languages. Even though CLIR and MLIR are tightly coupled, effective MLIR models need to overcome additional major challenges. For instance, realistically, the distribution of relevant documents across languages differs for each query. How can an MLIR model perform document ranking independent of its language to retrieve documents consistently? Methods for MLIR are categorized as (1) Fusion-based methods that first break MLIR into a group of CLIR sub-tasks and then merge results from multiple retrieval runs (Savoy, 2003; Savoy and Berger, 2005). From an IR perspective, the merging step depends on the assumption of the relevant distribution (Le Calvé and Savoy, 2000) or ranking score distribution (Manmatha et al., 2001), which often leads to sub-optimal results. (2) Direct methods that build the index for the entire collection and return one ranked list over multiple languages in the collection (Rahimi et al., 2015; Sorg and Cimiano, 2012). This line of research depends on different multilingual language modeling techniques. As previously discussed, the development of language models (e.g., mPLMs) often leads to language bias, later resulting in inconsistent retrieval performance across different languages.

In this dissertation, we developed MLIR models categorized under the direct method. These models fundamentally aim to handle and process multiple languages effectively within a single framework. To tackle language bias, Chapter 5 introduces a knowledge distillation approach by using English as the pivot language. This strategy involves transforming the representations from multiple languages into the English

[1] `https://en.wikipedia.org/wiki/Lists_of_languages`

embedding space, thus facilitating query document matching across languages. Chapter 6 advances this concept by targeting and eliminating language-dependent features from the representations of queries and documents in a dense retrieval model. This process reduces the influence of language that could skew retrieval outcomes and encourages retrieval to focus solely on content relevance. The primary objective of these methods is to refine the model's capability to rank documents by their relevance to the query, independent of the language in which they are in. This is achieved by ensuring that the core semantic content of the query and document guides the retrieval process rather than linguistic characteristics.

CHAPTER 3

TRANSLATION KNOWLEDGE INJECTION

Pre-trained language models (PLMs) offer big gains for many downstream tasks, including document ranking (Yates et al., 2021; Zhan et al., 2020). The multilingual versions of such models (mPLMs) provide the possibility of bypassing the translation step and jointly learning many languages within the same model. Although big gains are expected with such joint training, in the case of cross-lingual information retrieval (CLIR), the models in a multilingual setting are not achieving the same level of performance as those in a monolingual setting (Bonab et al., 2020; Ruder et al., 2019). Reviewing the pre-training task of mPLMs, we find that the token representations in different languages are generated by the context rather than their translations. This raises the same concerns of *translation gap* pointed out by Bonab et al. (2020) in the static word embeddings. When developing the neural ranking models, though semantic similarity signals can tackle term mismatch problems, the *exact* matching of terms in documents with those in queries is still the most important signal in ad-hoc retrieval (Guo et al., 2016). In the monolingual retrieval task, it is easier for the neural model to identify the query terms that occur in documents because of the same lexical form. However, in the cross-lingual setting, such co-occurrence becomes mutually translated words. Projecting words in different languages into the same hyperspace, mPLMs tend to "translate" query terms into related terms – i.e., terms that appear in a similar context – in addition to or sometimes rather than synonyms in the target language (Pires et al., 2019). This property creates difficulties for the

model in connecting terms in multiple languages that co-occur in both query and document.

To address this issue, we build a novel Mixed Attention Transformer (MAT) to incorporate external word-level translation knowledge, such as a dictionary or translation table. We design a hybrid architecture to embed MAT into the transformer-based deep neural models. By encoding the translation knowledge into an attention matrix, the model with MAT is able to focus on the mutually translated words in the input sequence. Experimental results demonstrate the effectiveness of the external knowledge and the significant improvement of MAT-embedded neural reranking model on CLIR task.

The work described in this chapter, namely Mixed Attention Transformer for Leveraging Word-Level Knowledge to Neural Cross-Lingual Information Retrieval, was published in CIKM 2021 (Huang et al., 2021). I was the lead author who designed the model architecture and conducted the experiments.

3.1 Word-Level Knowledge for Translation Attention

Our goal is to incorporate additional knowledge from external translation references into a transformer architecture to enable it to more accurately connect query and document tokens based on translations. We define translation reference as a large structural dataset containing knowledge to translate words from one language to another, e.g., a human-constructed dictionary or a translation table built from parallel corpora. Suppose there exists a word-level translation reference T. Given word w_s in the source language and w_t in the target language, $T(w_t, w_s)$ returns the probability of w_s being translated to w_t.

$$T(w_t, w_s) = P(w_t|w_s, T)$$

Algorithm 1: Generate translation attention matrix.

Input: $[q, d]$ and $T(\cdot, \cdot)$
Output: M^{tr}
1 Initialize M^{tr} as a $m \times m$ zero matrix.
2 **for** *each token* w_k *in the input sequence* **do**
3 | $M^{tr}_{kk} = 1$
4 **end**
5 **for** *each query token* w_i **do**
6 | **for** *each document token* w_j **do**
7 | | $M^{tr}_{ij} = M^{tr}_{ji} = T(w_j, w_i)$
8 | **end**
9 **end**
10 $M^{tr} \leftarrow \text{RowNorm}(M^{tr})$
return: M^{tr}

We assume the query is in the source language with length of m_q words and the document is in the target language with length of m_d words. Therefore, the concatenation of query and document $[q, d]$ has length $m = m_q + m_d + 2$ (including special token `[CLS]` and `SEP`). Then we construct an $m \times m$ translation attention matrix M^{tr} based on $[q, d]$ and $T(\cdot, \cdot)$ by symmetrically assigning translation probabilities between query tokens and document tokens. We provide detailed instructions for constructing M^{tr} in Algorithm 1.

Note that the k^{th} row of M^{tr} represents the attention weights of k^{th} token in the input assigned across all the input tokens. In Algorithm 1, lines 2-4 guarantee each token, including out-of-vocabulary words, is assigned a weight to itself and the self weight is the upper bound of all of its translation probabilities. If q_i and d_j are mutually translated words, they get their translation probabilities to each other from lines 5-9. Finally, the row normalization ensures that the attention weights for each input token sum up to 1.

To encode rare words with limited vocabulary size, pre-trained language models often use Byte Pair Encoding (BPE), which splits words into sub-word units. Evidence shows that self-attention treats split words differently than non-split ones (Correia

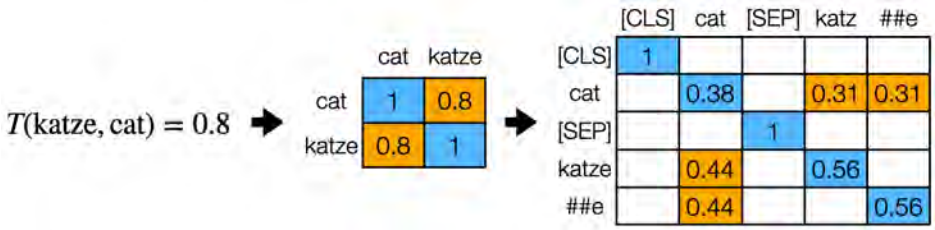


Figure 3.1: A simple example for generating M^{tr}.

et al., 2019). Therefore, we use tokens before BPE to query the translation reference and assign the same attention weight to all parts of the same word. The dimension m of M^{tr} is the same as the length of the $[q, d]$ sequence tokenized by a pre-trained language model. A simplified example for generating M^{tr} with query "cat" and document "katze" (German translation of cat) is shown in Figure 3.1.

3.2 Mixed Attention Transformers for CLIR

3.2.1 Architecture of Mixed Attention Transformers (MAT)

To inject M^{tr} into a transformer-based model, we design a novel transformer network named Mixed Attention Transformer (MAT) by combining multi-head attention with translation-based attention. The multi-head attention (Vaswani et al., 2017) is the core of the transformer architecture, which consists of n different attention heads. Given the vector representations as the hidden states $\mathbf{h}$, each head computes the dot-product attention:

$$\text{Attention}_i(\mathbf{h}) = \text{softmax}\Big(\frac{W_i^q\mathbf{h} \cdot W_i^k\mathbf{h}}{\sqrt{d/n}}\Big)W_i^\nu\mathbf{h}$$

where $\mathbf{h}$ is a d dimensional hidden vector for an input sequence. In BERT, the W_i^q, W_i^k and W_i^ν are matrices with size $d/n \times d$. Thus, each head projects to a different subspace of size d/n, learning different information.

Then the outputs of the multi-head attention, $\mathrm{MH}(\cdot)$, are n heads concatenated and linearly transformed:

$$\mathrm{MH}(\mathbf{h}) = W^o[\mathrm{Attention}_1, \ldots, \mathrm{Attention}_n]$$

In parallel to multi-head attention, we introduce the translation attention head denoted as $\mathrm{TH}(\cdot)$. Inspired by the scaled dot-product attention, we replace the attention weights learned from matrices W_i^q and W_i^k by the fixed attention weights in M^{tr}. Then, the multi-head attention becomes a single fixed attention head as follows

$$\mathrm{TH}(\mathbf{h}) = W^o_{\mathrm{TH}}\big(M^{tr}(W^\nu_{\mathrm{TH}}\mathbf{h})\big),$$

where both W^o_{TH} and W^ν_{TH} are trainable matrices in $\mathrm{TH}(\cdot)$ with dimension $d \times d$. By matrix multiplying with M^{tr}, the translation attention head is capable of reducing the distance between mutually translated tokens in the token representation hyperspace. Therefore, the attention matrix M^{tr} "pays attention" to all these pairs of words and $\mathrm{TH}(\cdot)$ tends to "pulls" their hidden representations closer in the hyperspace. In the following, we prove the effect of M^{tr} on hidden states.

Lemma 3.1 Let convex combinations of vectors A and B be $\alpha A + \beta B$ and $\beta A + \alpha B$ where $\alpha + \beta = 1$. Then, the cosine similarity between $\alpha A + \beta B$ and $\beta A + \alpha B$ is greater or equal to the cosine similarity between A and B.

Proof.

$$\begin{aligned} Sim(\alpha A + \beta B, \beta A + \alpha B) &= \frac{(\alpha A + \beta B) \cdot (\beta A + \alpha B)}{\|\alpha A + \beta B\| \|\beta A + \alpha B\|} \\ &\geq \frac{(\alpha^2 + \beta^2) A \cdot B + \alpha\beta(\|A\|^2 + \|B\|^2)}{(\alpha^2 + \beta^2)\|A\|\|B\| + \alpha\beta(\|A\|^2 + \|B\|^2)} \\ &\geq \frac{A \cdot B}{\|A\|\|B\|}. \end{aligned}$$

Therefore, $Sim(\alpha A + \beta B, \beta A + \alpha B) \geq Sim(A, B)$. □

Suppose query word w_i and document word w_j are the translations of each other with probability $p > 0$, and words other than w_j in documents all have zero translation probability with w_i. Then, the only two non-zero weights in the i^{th} row of M^{tr} are self attention (M^{tr}_{ii}) and attention on w_j (M^{tr}_{ij}):

$$M^{tr}_{ii} = \frac{1}{(1+p)};\ M^{tr}_{ij} = \frac{p}{(1+p)}$$

Similarly for w_j, the non-zero weights in the j^{th} row are $M^{tr}_{jj} = 1/(1+p)$ and $M^{tr}_{ji} = p/(1+p)$. If we ignore the trainable matrices in $\text{TH}(\cdot)$ and directly multiply M^{tr} with hidden states $\mathbf{h}$, the translation attention output of w_i and w_j are a convex combination of each other's hidden representations:

$$\text{TH}(\mathbf{h}_{w_i}) = \frac{1}{1+p}\mathbf{h}_{w_i} + \frac{p}{1+p}\mathbf{h}_{w_j}$$

$$\text{TH}(\mathbf{h}_{w_j}) = \frac{1}{1+p}\mathbf{h}_{w_j} + \frac{p}{1+p}\mathbf{h}_{w_i}$$

According to **Lemma 3.1**, because $p > 0$,

$$Sim\big(\text{TH}(\mathbf{h}_{w_i}), \text{TH}(\mathbf{h}_{w_j})\big) > Sim(\mathbf{h}_{w_i}, \mathbf{h}_{w_j})$$

Thus, when p is significant, the words in the query and document are likely to be translated to each other. The attention matrix M^{tr} "pays attention" to all these pairs of words and "pull" their hidden representations closer in the hyperspace.

The complete attention mechanism in MAT is a combination of the attention outputs from both $\text{MH}(\cdot)$ and $\text{TH}(\cdot)$. We first employ a residual connection around

each type of attention output, followed by layer normalization, denoted as LN(·), resulting two sub-layer outputs. Then we sum two sub-layer outputs:

$$\text{Sublayer}_{\text{MH}}(\mathbf{h}) = \text{LN}(\mathbf{h} + \text{MH}(\mathbf{h}))$$

$$\text{Sublayer}_{\text{TH}}(\mathbf{h}) = \text{LN}(\mathbf{h} + \text{TH}(\mathbf{h}))$$

$$\mathbf{h}' = \text{Sublayer}_{\text{MH}}(\mathbf{h}) + \text{Sublayer}_{\text{TH}}(\mathbf{h})$$

and apply the summed result to the position-wise feed-forward networks (FFN):

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

The final output of MAT is another residual connection around the output of FFN:

$$\text{MAT}(\mathbf{h}) = \text{LN}(\mathbf{h}' + \text{FFN}(\mathbf{h}'))$$

The complete MAT architecture is depicted in Figure 3.2 (middle). The left and right of Figure 3.2 are two types of attention components in MAT. The benefits of this network architecture are that the MAT can attend to both contextual information from multi-head attention and cross-lingual knowledge from the translation attention head during training. Because we keep the multi-head attention mechanism and share the FFN sublayer, MAT contains a vanilla transformer network. This design allows MAT to be easily embedded into recent transformer-based pre-trained models and fully leverage the pre-trained weights.

3.2.2 Embed MAT into Pre-trained Language Models

Qiao et al. (2019) analyzed different ranking models based on BERT and found that the cross-encoder approach, which applies BERT on the concatenated $[q, d]$ sequence and uses the last layer's representation of the `[CLS]` token as the matching

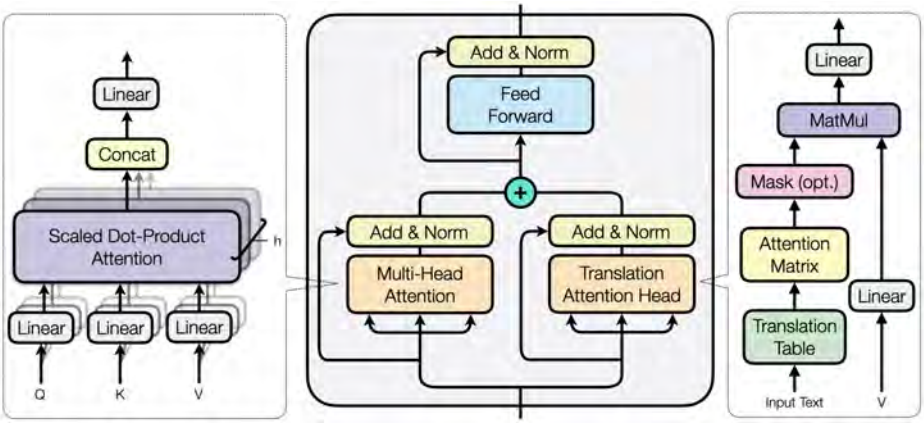


Figure 3.2: (left) Multi-Head Attention. (right) Translation Attention Head. (middle) Mixed Attention Transformer Layer.

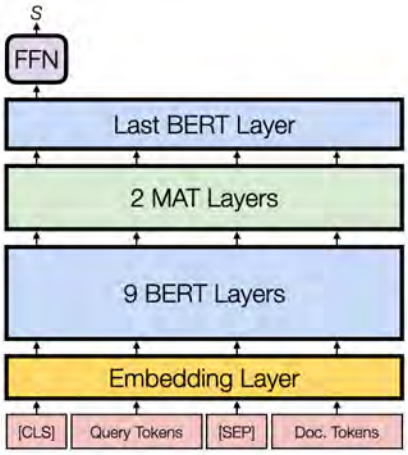


Figure 3.3: MAT layers in BERT document ranking model.

feature, gives the best performance. We use the same BERT in the cross-encoder setting as a re-ranker to discuss how to embed MAT into a transformer-based pre-trained language model.

MART (**MA**T+BE**RT**), the new model architecture we proposed, keeps the embedding layer and add-on network while replacing some of the Transformer layers in the middle by MAT. During fine-tuning, the BERT layers close to the output (higher layers) are more sensitive than the lower layers (Zhao and Bethard, 2020). Also, another study on BERT (Tenney et al., 2019) has shown that most local syntactic

phenomena are encoded in lower layers while higher layers capture more complex semantics. Considering the fine-tuning efficiency and semantic quality of the token representations, the layer replacement should start from the higher layers of BERT. Moreover, in the `Last-Int` ranking approach, the output score is only based on the `[CLS]` token in the last BERT layer. Therefore, we keep the last BERT (`Base`) layer as the output layer and start to embed MAT from the 11^{th} layer. Figure 3.3 shows an example of the hybrid architecture based on a BERT-based ranking model where MAT layers are embedded into 10^{th} and 11^{th} layers of BERT. Using the same hidden dimension as BERT, each MAT layer introduces only about 1.18M new parameters compared to the BERT layer. At initialization, MAT is able to use pre-trained weights of its corresponding BERT layer. This compatibility increases the fine-tuning efficiency and reduces the training data requirement.

3.3 Experimental Setup

3.3.1 Datasets

CLIR Dataset. We create our training and evaluation data from the Cross-Language Evaluation Forum (CLEF) 2000-2008 campaign for bilingual ad-hoc retrieval tracks (Braschler, 2001, 2002a,b, 2003; Peters, 2005, 2006, 2007, 2008, 2009). We use the text fields of the documents to construct our retrieval corpus and discard other metadata. We concatenate the title and description fields of a topic and consider it as our query. We consider all the topics and relevance judgments from all the tracks to show the consistent effectiveness of MAT across several cross-language retrieval settings on both high- and low-resource languages.

Translation Resources. Our goal is to leverage translation resources as external knowledge into the query-document matching process. We use sentence-level parallel data with GIZA++ toolkit (Och and Ney, 2003) to construct a translation table, which we use to generate M^{tr}. Translation tables for European languages are based

on the Europarl v7 sentence-aligned corpora (Koehn, 2005). For our limited-resource (in terms of both parallel data and relevance judgments) setting based on Somali and Swahili languages, we use the translation tables provided by Zhang et al. (2020).

Forward Setting: Non-English Query and English Documents. In this setting, we use non-English queries against an English document collection. To evaluate cross-lingual matching performance, we use human translation of a fixed query set to obtain queries in different languages. While we have translations of queries in different languages, we keep the content and language of the retrieval corpus fixed. We have both high-resource and low-resource CLIR settings in our experiments. In a high-resource setting, for example, French-English, we have a larger size of sentence-level parallel data and relevance judgments compared to a low-resource setting. In our experiments, we used four high-resource language pairs: French (Fre-Eng), Italian (Ita-Eng), German (Deu-Eng), and Spanish (Spa-Eng). For each language, we selected queries from the CLEF C001 – C350 topic set. We took the intersection of the topic ID and removed topics without any relevant documents, resulting in 246 overlapped queries across four languages. For cross-language information retrieval involving low-resource languages, we experiment on Somali (Som-Eng) and Swahili (Swa-Eng). Bonab et al. (2019) provided Somali and Swahili translations of 151 English queries from the CLEF C001 – C200 topic set, and we use those queries in our setting. The collection of English documents is the Los Angeles Times corpus comprised of 113K news articles.

Backward Setting: English Query and Non-English Documents In this setting, we use English queries to search document collections in four languages: French (Eng-Fre), Italian (Eng-Ita), German (Eng-Deu) and Spanish (Eng-Spa). For each language, we create a retrieval corpus from a combination of sources which we report in Table 3.1. As the retrieval corpus varies for each language, relevance judgments are not available for all the English topics from CLEF C001 – C350 topic set.

Table 3.1: Summary of CLIR setting. The first four rows indicate the backward and the last row indicates the forward setting.

CLIR Setting	Collection Source	Collection Size	Query Size
Eng-Fre	Le Monde, Sda French	129,689	185
Eng-Ita	La Stampa, Sda Italian	144,040	176
Eng-Deu	Der Spiegel, Frankfurter Rundschau	153,496	184
Eng-Spa	EFE News 94-95	452,027	156
Xxx-Eng	Los Angeles Times 94	113,005	246

Thus, for each CLIR setting, we have a different number of queries in the backward setting compared to the forward setting. Table 3.1 provides information about query sets and document collections in both settings.

3.3.2 Implementation Details

Passage Re-ranker. Nogueira and Cho (2019) fine-tuned mBERT on MS MARCO passage retrieval dataset to create a passage ranking model. We refer to this model checkpoint as m^{2}BERT and further fine-tune it with cross-lingual relevance judgments. To prepare the input sequence for m^{2}BERT we concatenate a query and a document separated by a special `[SEP]` token from mBERT's vocabulary. We prefix the concatenated sequence with the special `[CLS]` token from mBERT's vocabulary. We obtain the last layer representation of this sequence from m^{2}BERT, but only use the representation of the `[CLS]` token, and pass it through a linear combination layer to obtain the probability of the document being relevant to the query. At test time, given a query, m^{2}BERT computes the probability for each document independently and obtains a document ranking after sorting with these probability scores. Because the mBERT input sequence is limited to 512 tokens, longer documents are split into segments by 512 tokens, and `[CLS]` representations from all document segments are averaged to obtain a representation for fine-tuning. MacAvaney et al. (2019) used the same approach for monolingual retrieval.

Evaluation. For evaluating retrieval effectiveness, we follow prior works on the CLEF dataset (Bonab et al., 2020; Litschko et al., 2019) and report mean Average Precision (mAP) of the top 100 ranked documents and precision of the top 10 retrieved documents (P@10). We determine statistical significance using the two-tailed paired t-test with p-value less than 0.05 (i.e., 95% confidence level).

Model Training. We train all neural re-ranking models using pairwise cross-entropy loss (Dehghani et al., 2017). We use all the positive documents from the query relevance judgments and randomly sample negative documents to form training pairs. All models are trained using Adam's optimization algorithm (Kingma and Ba, 2015) with a learning rate of 2e-5, batch size 16, and 100 epochs with an early stopping strategy. Given the limited number of queries in each language, we use 5-fold cross-validation for robust evaluation. For each fold, the training, validation, and test data are 60%, 20%, and 20% of the query set, respectively. The reported evaluation metrics are averaged across 5 folds. We also fix the random seed to guarantee that all models receive the same training data. For the validation queries, we re-rank the top 100 documents and use mAP to select the best-performing model.

3.3.3 Baselines

We compare MART with the methods in the following

- **SMT+BM25**: We first use the GIZA++ toolkit (Och and Ney, 2003) to build translation tables from parallel corpora. We select top-10 translations from the translation table for each query term and apply Galago's[1] weighted *#combine* operator to form a translated query. Then we use the Galago's implementation of Okapi BM25 (Robertson et al., 1995) with default parameters. It serves as one of our baselines. Moreover, the training data for neural re-ranking models are sampled based on the top 500 retrieved documents by SMT+BM25.

[1] `https://www.lemurproject.org/galago.php/`

- **NMT+BM25**: Leveraging OPUS-MT (Tiedemann and Thottingal, 2020), an open-source NMT project for many languages, we build this baseline by first translating the query into the same language as the documents using an NMT model. Then, we run BM25 to retrieve documents.
- **m^{2}BERT**: To create the m^{2}BERT baseline, we begin with the pre-trained checkpoint provided by Nogueira and Cho (2019). This checkpoint is a result of fine-tuning the multilingual BERT (mBERT) architecture with MS MARCO passage ranking dataset (Nguyen et al., 2016). We further fine-tune it with training data from a specific CLIR setting. We use the same fine-tuning approach described in section 3.3.2 for this baseline and our proposed model to ensure a fair comparison.
- **MART-PLB**: This is a variant of MART. In order to evaluate the effect of external knowledge on MAT, we replace M_{tr} by an identity matrix so that each token is only paying attention to itself. Therefore, instead of injecting translation knowledge into the model, we design a "placebo" attention matrix for MAT. Using MART-PLB as a controlled experiment, we are able to evaluate the effect of external knowledge.

In order to provide an empirical upper bound on retrieval performance, we use human translation of the queries and apply BM25 as the retrieval technique, denoted as **Human+BM25**. The human translations of the queries are obtained from the CLEF dataset, as they have a common topic ID for the same queries across different languages.

3.4 Results

3.4.1 Performance on High-resource Languages

Table 3.2 lists evaluation results on both Forward (top) and Backward (bottom) settings for language pairs with high translation resources. Based on BM25, the

Table 3.2: Model performance on forward and backward settings for high-resource languages. The highest value for each column is marked with bold text. Statistically significant improvements are marked by † (over SMT+BM25), ‡ (over NMT+BM25) and ⋆ (over BERT).

	Model	Fre-Eng		Ita-Eng		Deu-Eng		Spa-Eng	
		mAP	P@10	mAP	P@10	mAP	P@10	mAP	P@10
English Docs (Forward)	Human+BM25	0.4569	0.3940	0.4569	0.3940	0.4569	0.3940	0.4569	0.3940
	SMT+BM25	0.3618	0.3492	0.3561	0.3431	0.3588	0.3354	0.3624	0.3317
	NMT+BM25	0.4029	0.3682	0.3637	0.3282	0.3586	0.3306	0.3623	0.3310
	m^{2}BERT	0.3802†	0.3799†	0.3652	0.3545	0.3582	0.3335	0.3819†	0.3693†
	MART-PLB	0.3859†	0.3666†	0.3701	0.3689†	0.3593	0.3501†	0.3824†	0.3676†
	MART	**0.4126**†*	**0.3935**†‡*	**0.3944**†‡*	**0.3732**†‡*	**0.3862**†‡*	**0.3770**†‡*	**0.3953**†‡*	**0.3830**†‡*
	Model	Eng-Fre		Eng-Ita		Eng-Deu		Eng-Spa	
		mAP	P@10	mAP	P@10	mAP	P@10	mAP	P@10
English Queries (Backward)	Human+BM25	0.2955	0.3054	0.2629	0.2892	0.2970	0.3060	0.2518	0.2436
	SMT+BM25	0.2258	0.2319	0.1883	0.1852	0.2614	0.2424	0.1985	0.2088
	NMT+BM25	0.2397	0.2584	0.1943	0.2119	0.2603	0.2734	0.2477	**0.3532**
	m^{2}BERT	0.2841†	0.2875†	0.2635†	0.2605†	0.3241†	0.3246†	0.2355†	0.2285†
	MART-PLB	0.2807†	0.2823†	0.2713†	0.2771†	0.3262†	0.3230†	0.2389†	0.2351†
	MART	**0.3002**†‡*	**0.3108**†‡*	**0.2823**†‡*	**0.2846**†‡*	**0.3433**†‡*	**0.3414**†‡*	**0.2558**†*	0.2439†*

difference between NMT and SMT is due to translation quality. We can see that NMT outperforms SMT on both forward and backward translation of French and Italian tasks. They perform similarly on German tasks. In Spanish, NMT models show a different translation quality between forward and backward.

As a neural re-ranker, m^{2}BERT significantly improves upon SMT+BM25 on all language pairs in the backward setting (English queries) and two language pairs in the forward setting (English docs) while performing on par with SMT+BM25 for Deu-Eng and Ita-Eng languages. While fine-tuned on English document retrieval dataset, m^{2}BERT can transfer to the cross-lingual task with small amount of fine-tuning data. This agrees with the previous finding by Pires et al. (2019) that mBERT is able to generalize across languages.

We observed substantial improvements on the retrieval performance when translation knowledge is incorporated into MART. For all language setting combina-

tion in Table 3.2, MART performs significantly better than the BERT architecture (m^{2}BERT) in terms of both mAP and P@10. MART improves m^{2}BERT by 8% on the forward and 7% on the backward settings in terms of mAP. This comprehensive comparisons with vanilla BERT based ranker demonstrate the effectiveness of the MAT-embedded model.

Replacing M_{tr} by the identity matrix in MART-PLB, the translation attention head degenerates to two additional feed-forward layers. MART-PLB behaves insignificantly different comparing to the vanilla BERT architecture on all languages. Such results indicate that the performance gain in MART relies on injecting the external knowledge, not from adding new parameters. When M_{tr} becomes non-informative, the translation attention head is ineffectual.

Comparing MART with human translation, we can see that in the forward setting, correct translation with the basic retrieval model still beats the neural CLIR model. However, in the backward setting, MART achieves roughly the same as (Eng-Fre, Eng-Spa) or better than (Eng-Ita, Eng-Deu) human translation. We hypothesize that in the backward setting, translation tables provide higher-quality translations, which enable better semantic matching between query and document tokens.

3.4.2 Performance on Low-resource Languages

The evaluation results for two language pairs with limited translation resources on the forward setting are shown in Table 3.3. We make several observations. First, different from high-resource languages, we can see that, on the Somali-English task, NMT performs worse than SMT due to the limited resources for NMT model training. Then, m^{2}BERT mostly underperforms SMT+BM25 for both Somali and Swahili languages. Note that Somali is not included in the mBERT pre-training. Even if Swahili is included, there are only a small number of Swahili sentences in the pre-training data. The low performance of m^{2}BERT on low-resource language pairs demonstrates

Table 3.3: Model performance for low-resource languages on Forward setting. The highest value for each column is marked with bold text. Statistically significant improvements are marked by † (over SMT+BM25), ‡ (over NMT+BM25) and ⋆ (over m^{2}BERT).

Model	Som-Eng		Swa-Eng	
	mAP	P@10	mAP	P@10
Human Translation	0.4563	0.3940	0.4563	0.3940
SMT+BM25	0.1948	0.1865	0.2184	**0.2152**
NMT+BM25	0.1589	0.1380	0.2247	0.2113
m^{2}BERT	0.1986	0.1772	0.2055	0.2089
MART-PLB	0.2049	0.1972†*	0.2130	0.2106
MART	**0.2207**†‡*	**0.2135**†‡*	**0.2348**†*	0.2151

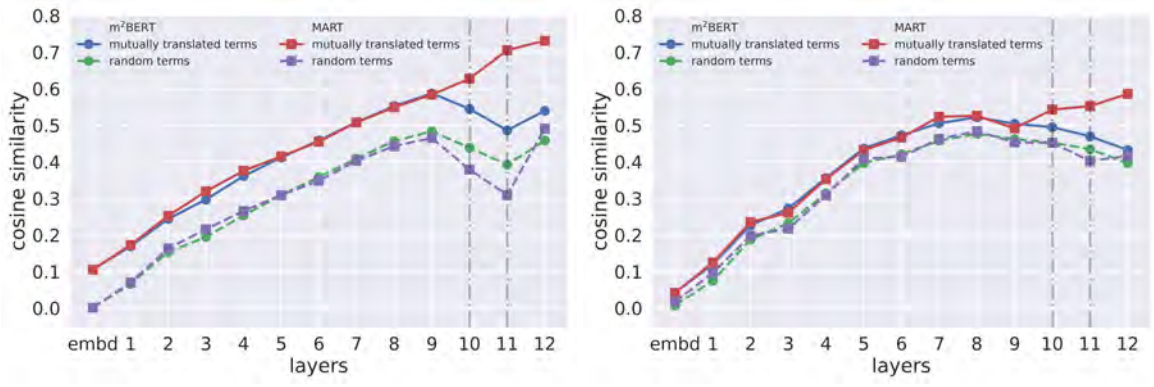
that absence or inadequate pre-training data on a particular language leads to poor performance on target tasks involving those languages.

On the other hand, the MART model achieves the highest mAP performance for both Somali and Swahili languages. The consistent and significant improvements in terms of mAP over compared methods make MART the best model in our experiments. Due to the lack of pre-training data, the translation gap is more critical in low-resource language pairs. The performance of MART for Somali and Swahili languages proves that leveraging external translation knowledge can help to bridge the translation gap. Moreover, the experiments with the placebo setting, similar to those for the high-resource languages, have shown no significance in performance compared to m^{2}BERT. These results strengthen the conclusion that the translation attention matrix is the key component of MAT.

Human translation leads neural ranking models by a large margin in CLIR tasks involving low-resource languages. This is expected because, with less sentence-level parallel data, the CLIR models often suffer from low quality of translations.

3.4.3 Analysis of Token Representations

To study the influence of MAT on the translation gap in neural CLIR, we compare the token representation from each layer between m^{2}BERT and MART. Specifically,



(a) Deu-Eng (b) Swa-Eng

Figure 3.4: The comparison of layer-wise token representations.

both models are fine-tuned on Deu-Eng and Swa-Eng training data. Figure 3.4 shows the distances between contextualized token representations in two model architectures where the x-axis represents layers from low to high and the y-axis is the cosine similarity. We focus on two types of word pairs (one from query and another from document) in an input sequence: (i) Mutually translated words, where all pairs of words that are translations of each other according to the external translation knowledge are selected; and (ii) Random non-translated words, where we randomly sample 10 pairs of words which are not translations of each other. We compute the average cosine similarity of the token representations at each layer for all selected word pairs in the test data of Deu-Eng (high-resource) and Swa-Eng (low-resource).

From the diagrams in Figure 3.4, we can see that, in general, the similarity of token representations increases as the layer gets higher. Also, mutually translated words always have smaller cosine distances than non-translated words. The closer lines between two types of word pairs in Swa-Eng prove that the translation gap is more critical in resource-lean languages. We can also see that in the 10^{th} and 11^{th} layers, the similarity of two types of words in m^{2}BERT drops for both language pairs. According to the previous analysis (Pires et al., 2019), one hypothesis for such drop is that before fine-tuning on MS MARCO dataset, mBERT was pre-trained on

surrounding contexts for language modeling, it needs more contextual information to predict the missing words correctly. Therefore, mBERT favors text sequence pairs that are closer in their semantic meanings. Such models trained on surrounding context are not as effective for ad-hoc document ranking with respect to keyword queries (Qiao et al., 2019).

MART shows the same behavior as m^{2}BERT up to the MAT layers. The representations of mutually translated words in MAT layers become similar to each other in terms of cosine distance. This matches the design purpose of MAT. Meanwhile, because MAT keeps the native multi-head attention from BERT layer, the similarity of non-translations still drops in MAT layers. The increased similarity between mutually translated words and the decreased similarity between non-translated words demonstrate that the model is bridging the translation gap with the help of external knowledge.

3.4.4 Effect of Translation Resources

From the previous results, we have seen that the translation attention matrix is critical to the success of MAT. As a knowledge injection model, it is palpable that the quality of the knowledge affects the model performance. In this experiment, we study the effect of different sources of external knowledge on the MART. Besides the translation table built from parallel data, we use two different translation knowledge for M_{tr} generation: Panlex dictionary (Kamholz et al., 2014) and multilingual word embedding, MUSE (Conneau et al., 2017). To obtain translation probability for a single word in Panlex, we uniformly distribute weights to all possible translations. In MUSE, we use the five nearest neighbors of a word in the target languages as its potential translations and assign translation probability based on their normalized cosine similarity. In order to cover different languages and retrieval settings, we

Table 3.4: MART performance for different external knowledge. The highest value for each column is marked with bold text. "–" if language is not supported.

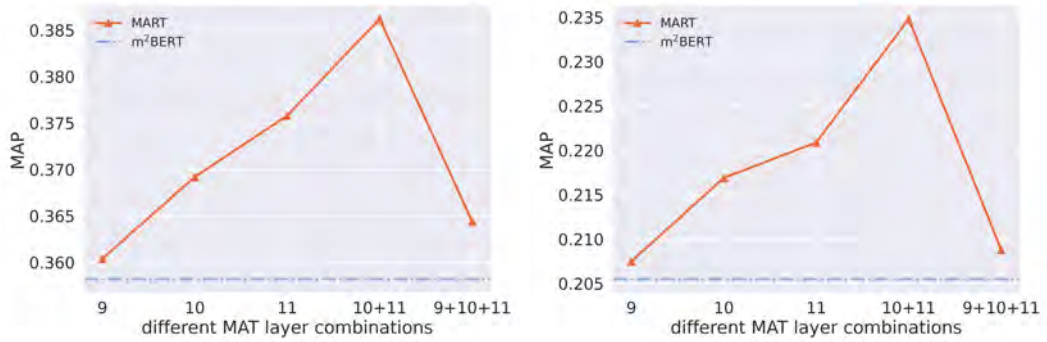
External Knowledge	Forward				Backward	
	Deu-Eng		Swa-Eng		Eng-Deu	
	mAP	P@10	mAP	P@10	mAP	P@10
Parallel Corpus	**0.3862**	**0.3770**	**0.2348**	**0.2151**	**0.3433**	**0.3414**
Panlex	0.3713	0.3612	0.2265	0.2073	0.3326	0.3360
MUSE	0.3693	0.3580	–	–	0.3335	0.3348

select Deu-Eng (high-resource) and Swa-Eng (low-resource) from forward setting and Eng-Deu from backward setting for this experiment.

Table 3.4 shows the results of all compared sources of translation knowledge. We observe a performance drop on both alternative knowledge resources. For Panlex, although the translations are more precise than those in a translation table, they do not provide a broad coverage of words. Multilingual word embeddings are learnt from the contexts of words, not their translations. Therefore, given a word, the embeddings of semantically similar words are often closer than those of its translations to the embedding of a word (Bonab et al., 2020). Thus, using multilingual word embeddings, the problem of the translation gap will not be completely resolved.

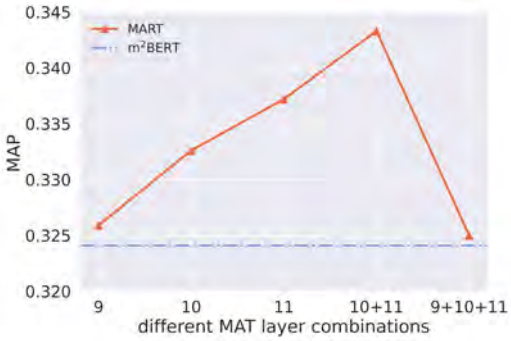
3.4.5 Ablation Study on Model Architecture

In this section, we empirically study the effects of different numbers and positions of MAT layers in a MART model. We further train and evaluate the MART with various combinations of MAT layers. It is worth mentioning that given the number of layers in BERT architecture, there exist exponential number of possible combinations. We only explore several representative models. Leaving the last layer as the output layer, we still focus on the higher Transformer layers of BERT architecture. For models with a single MAT layer, we investigate MART with MAT embedded at 9^{th}, 10^{th}, or 11^{th} layer. For double MAT layers, we use the previous results from MAT at



(a) Deu-Eng

(b) Swa-Eng



(c) Eng-Deu

Figure 3.5: The performance comparison of different MART model architectures.

10^{th} and 11^{th} layers. We also consider an architecture with three MAT layers where 9^{th}, 10^{th} and 11^{th} layers in BERT are all replaced by the MAT layer.

Figure 3.5 shows the performance of different MART model architectures on Deu-Eng, Swa-Eng and Eng-Deu. We can see that all model variants have a similar pattern across three selected CLIR tasks. Because higher BERT layers are more sensitive to fine-tuning (Zhao and Bethard, 2020) and their hidden representations capture complex semantic information (Tenney et al., 2019), the retrieval performance for the single MAT layer increases from MAT at the 9^{th} layer to MAT at the 11^{th} layer. The double MAT layer can further boost performance from the single-layer approach. We also can see that models get less improved when 9^{th} in replaced by MAT. This pattern indicates that the token representations after the 8^{th} layer (the

input of the 9^{th} layer) do not contain enough semantic information. Thus, it is too early to apply the translation attention.

3.5 Summary

In this chapter, we bridged the translation gap in neural CLIR models by incorporating external translation knowledge. First, we build an attention matrix for mutually translated words between query and document based on the translation resource. Then using the attention matrix, we design a new translation attention head and show that it is able to reduce the cosine distance between hidden representations of mutually translated words. Finally, the complete architecture of MAT is a combination of multi-head attention and translation attention head with shared feed-forward networks. As a layer component, we further design an architecture to embed MAT into Transformer-based CLIR models. Our comprehensive experimental results demonstrate the effectiveness of external knowledge and the significant improvement of the MAT-embedded neural model on the CLIR task.

However, we can see that in this approach, the cross-lingual knowledge is extracted as the static word translations prior to model training. And tuning of the model parameters still relies on the cross-lingual retrieval data. The lack of high-quality retrieval data often limits the performance of models in CLIR tasks. In the next chapter, we will investigate cross-lingual knowledge transfer to construct CLIR models without requiring retrieval labels between two languages.

CHAPTER 4

CROSS-LINGUAL TRANSFER VIA KNOWLEDGE DISTILLATION

In this chapter, we focus on *retrieval data scarcity*, another challenge that prevents CLIR models from achieving similar performance as the monolingual (e.g., English) retrieval models. PLM-based dense retrieval models, such as DPR (Karpukhin et al., 2020) and ColBERT (Khattab and Zaharia, 2020), have shown promising performance on English retrieval tasks. This success is mainly due to two key factors: (i) The unsupervised pre-training of context-aware transformer architectures with an enormous number of parameters over large corpora. (ii) The fine-tuning for the downstream learning-to-rank task with a relatively large collection of relevance judgments, such as the MS MARCO passage ranking dataset (Nguyen et al., 2016).

Replacing the backbone encoder from PLMs with mPLMs (multilingual PLMs) allows for the joint learning of many languages with the same model. Because mPLMs project tokens in different languages into the same space, fine-tuning these models with a cross-lingual retrieval dataset, similar to the monolingual setting, enables cross-language information retrieval. However, both factors leading to the success of monolingual information retrieval have defects in the cross-lingual setting. First, due to the unbalanced pre-training data in different languages, mPLMs have already shown a performance gap between high and low-resource languages in many downstream tasks (Wang et al., 2020; Wu and Dredze, 2020). Cross-lingual retrieval models built on such mPLMs can inherit the language bias, leading to suboptimal results for low-resource languages. Second, compared with the English-to-English retrieval task, the lack of cross-lingual training data with reliable relevance judgment, i.e., human

relevance judgments, especially for low-resource languages, makes it more challenging to learn cross-lingual retrieval models.

Studies have attempted to address the data scarcity problem in CLIR. Sasaki et al. (2018) proposed a large cross-lingual retrieval collection, WikiCLIR, based on the linked foreign language articles from Wikipedia pages. Because Wikipedia articles in a specific language are edited mainly by native speakers, the cross-lingual contents in WikiCLIR are of high quality. But the relevance judgments are synthetically generated based on mutual links across pages. On the other hand, Bonifacio et al. (2021) built a multilingual passage ranking dataset, mMARCO, by translating the queries and passages in MS MARCO into the target language using Neural Machine Translation (NMT) models. Because MS MARCO is generated from query logs, the relevance judgments in mMARCO are more credible than WikiCLIR. Still, the automatically generated cross-lingual content created by NMT models is not comparable to human writers, especially for resource-lean languages.

Different from training monolingual retrieval models, the cross-lingual retrieval data serves two purposes during the model training: (i) the languages of the query and document provide cross-lingual knowledge, and (ii) the relevance label of the query and document provide retrieval knowledge. In the previous chapter, cross-lingual knowledge is first extracted from a parallel corpus and stored as translation tables for models to inquire. In this chapter, we separate the learning of retrieval knowledge from the learning of cross-lingual knowledge and use the semantic closeness of bitext data to transfer an English retrieval model to the target language. Specifically, we present OPTICAL: Optimal Transport distillation for Cross-lingual information retrieval. Following the dense retrieval paradigm, we first train a bi-encoder English-to-English retrieval model, similar to the ColBERT architecture (Khattab and Zaharia, 2020), as the *teacher model*. Suppose the CLIR task is to search English documents with non-English queries. To devise a complete *student model* for this

CLIR task, we then reuse the teacher model's document encoder and train a new student query encoder for non-English queries. After training, given parallel queries, the non-English token representations generated by the student's query encoder should be similar to the English token representations generated by the teacher's query encoder.

The student model distills the retrieval knowledge from the teacher model in a cross-lingual setup. We form the distillation training as an optimal transport problem (Peyré et al., 2019) where the cost matrix is the cross-lingual token cosine distance. The optimal transportation plan serves as a soft token alignment. Since the teacher model has already learned query-document matching, the distillation training can concentrate solely on cross-lingual knowledge as a general text encoder. As a result, we can employ bitext data, a collection of parallel and comparable documents, to train the student query encoder. Compared to cross-lingual relevance labeling, bitext data can be mined by automatic algorithms (Heffernan et al., 2022), making it more practical for languages with limited resources.

The work described in this chapter, namely Improving Cross-lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation, was published in WSDM 2023 (Huang et al., 2023a). I was the lead author who designed the model architecture and conducted the experiments.

4.1 Cross-lingual Knowledge Distillation

Our goal is to incorporate the knowledge of query document matching from a well-learned monolingual retrieval model into a multilingual transformer-based retrieval architecture, such that it is capable of generating contextual representations under the cross-lingual setting and thus performing query document matching in different languages. In this section, we first introduce the monolingual retrieval model known as the teacher model. Then we present the optimal transport knowledge distillation framework and training of the student model. Finally, we combine the components

from both the teacher and student models to fulfill a CLIR task. To better describe our approach, we focus on the CLIR case of searching an English collection with a non-English query as an example to describe our method.

4.1.1 Teacher Model

The teacher model M contains two components: query encoder E_{M_q} and document encoder E_{M_d}. Given a query q and a candidate document d, the score of matching between q and d, $S_{q,d}$, is then computed as the:

$$S_{q,d} = \sum_{i=1}^{|q|} \max_{j=1}^{|d|} E_{M_q}(q_i) \cdot E_{M_d}^T(d_j) \tag{4.1}$$

where $E_{M_q}(q_i)$ is the i-th token representation of the query and $E_{M_d}(d_j)$ is the j-th token representation of the document. The scoring function applies the *maxsim* operation on each query token to softly search against all document tokens to find the best token that reflects its context and then sums over all the query tokens. The primary goal of the teacher model is to provide knowledge of query document matching regardless of the language. Therefore, we select the English MS MARCO passage ranking dataset for teacher model training because of its size. Similar to ColBERT, we prepend special tokens `[Q]` and `[D]` to the query and passage tokenization, respectively, and expand the query to a fixed length L using the `[MASK]` token. Unlike ColBERT, we initialize the teacher model using mBERT, instead of BERT. Since mBERT has a larger vocabulary that covers a more diverse set of languages, the student can benefit from the multilingual pre-trained token representation.

4.1.2 Optimal Transport Knowledge Distillation

The student model shares the same architecture as the teacher. Note that if the document collection for the CLIR task remains in English, then the document encoder of the student model E_{S_d} can be a copy of the teacher's document encoder,

E_{M_d}. Here, we focus on the design of the query encoder of the student model, E_{S_q}, which handles non-English queries. Assume that q is an English query and $\hat{q}$ is a non-English parallel query. The token representation of q generated by E_{M_q} contains rich knowledge for query document matching. If we could let E_{S_q} "behave" like E_{M_q}, namely, if the output of E_{S_q} with $\hat{q}$ is close to the output of E_{M_q} with q, then the token representations generated by E_{S_q} can have a similar retrieval performance to the teacher model. Therefore, the training objective of knowledge distillation is to reduce the distance between the outputs of the teacher and student query encoders given parallel inputs (sentences). Next, we define the distance from $\hat{q}$ to q in the vector space of E_{S_q} and E_{M_q}. Suppose the tokenizer tokenizes q into L_q tokens and $\hat{q}$ into $L_{\hat{q}}$ tokens. We expand them to the same length L by appending `[MASK]` tokens. After encoding by E_{S_q} and E_{M_q}, $\hat{q}$ and q are represented by a bag of vectors of size L, respectively. We define the distance from $\hat{q}$ to q as follows:

$$D(\hat{q}, q) = \underset{f_{[\![1..L]\!] \to [\![1..L]\!]}}{\arg\min} \frac{1}{L} \sum_{i=1}^{L} 1 - E_{S_q}(\hat{q}_i) \cdot E_{M_q}^T(q_{f(i)}) \tag{4.2}$$

where f is a bijective (one-to-one correspondence) function which maps the token index from $\hat{q}$ to q. Intuitively, the distance definition is equivalent to finding a token mapping from $\hat{q}$ to q that minimizes the average cosine distance among L token pairs. Despite the same semantics of $\hat{q}$ and q as a whole, different languages have different token arrangements. When L increases, using brute force to find the mapping f becomes computationally intractable. Therefore, we approximate the calculation of $D(\hat{q}, q)$ as an optimal transport problem (Peyré et al., 2019). First, we assign equal mass to the tokens in $\hat{q}$ and q by defining a uniform source probability distribution, μ_s, on $\hat{q}$ and a uniform target probability distribution, μ_t, on q: $\mu_s(i) = \frac{1}{L}$ and $\mu_t(j) = \frac{1}{L}$ where $1 \leq i, j \leq L$.

The set of transportation plans between these two distributions is then the set of doubly stochastic matrices $\mathcal{P}$ defined as

$$\mathcal{P} = \{\boldsymbol{\gamma} \in (\mathbb{R}^+)^{L \times L} \mid \boldsymbol{\gamma}\mathbf{1}_L = \mu_s,\ \boldsymbol{\gamma}^T\mathbf{1}_L = \mu_t\} \tag{4.3}$$

where $\mathbf{1}_L$ is a L-dimensional vector of ones and $\boldsymbol{\gamma}$ is called a transportation plan. We redefine the distance between $\hat{q}$ and q as a Wasserstein (earth mover's) distance between distribution μ_s and μ_t. Then the computation of such distance become an optimal transport (OT) problem:

$$\boldsymbol{\gamma}_0 = \arg\min_{\boldsymbol{\gamma} \in \mathcal{P}} \big\langle \boldsymbol{\gamma}, \mathbf{C} \big\rangle_F \tag{4.4}$$

where $\big\langle \cdot, \cdot \big\rangle_F$ is the Frobenius inner product, $\boldsymbol{\gamma}_0$ is the optimal transportation plan (or OT matrix) and $\mathbf{C} \geq 0$ is a $L \times L$ cost function matrix of term $C(i, j)$, reflecting the "energy" needed to move a probability mass from $\hat{q}_i$ to q_j. In our setting, this cost is chosen as the cosine distance between two tokens:

$$C(i, j) = 1 - E_{S_q}(\hat{q}_i) \cdot E_{M_q}^T(q_j)$$

In general, the linear programming solution to find $\boldsymbol{\gamma}_0$ has a typical super $O(n^3)$ complexity that is still computationally intractable (Cuturi, 2013). To overcome such intractability, we employ the Inexact Proximal point method for Optimal Transport (IPOT) (Xie et al., 2020) algorithm to compute the OT matrix. Specifically, the IPOT algorithm iteratively solves the OT problem by adding a proximity metric term to the original OT definition. At step t, we have:

$$\boldsymbol{\gamma}^{(t+1)} = \arg\min_{\boldsymbol{\gamma} \in \mathcal{P}} \Big\{ \big\langle \boldsymbol{\gamma}, \mathbf{C} \big\rangle_F + \beta \cdot \mathcal{B}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^{(t)}) \Big\}$$

where $\mathcal{B}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^{(t)}) = \sum_{i,j}^{L} \boldsymbol{\gamma}_{ij} \log \frac{\boldsymbol{\gamma}_{ij}}{\boldsymbol{\gamma}_{ij}^{(t)}} - \sum_{i,j}^{L} \boldsymbol{\gamma}_{ij} + \sum_{i,j}^{L} \boldsymbol{\gamma}_{ij}^{(t)}$ is the Bregman divergence used as a proximity metric term to penalize the distance between the solution and the latest approximation. It provides a tractable iterative scheme toward the exact

Algorithm 2: Inexact proximal point method for optimal transport (IPOT).

Input: Probability mass function of source and target $\{\mu_s, \mu_t\}$, cost matrix $\mathbf{C}$ and step size β.
Output: Approximated OT matrix $\widetilde{\boldsymbol{\gamma}}$

Function `IPOT`$(\mu_s, \mu_t, \mathbf{C}, \beta)$**:**
  $\boldsymbol{b} \leftarrow \frac{1}{L}\mathbf{1}_L$, $\boldsymbol{\gamma}^{(1)} \leftarrow \mathbf{1}\mathbf{1}^T$
  $\mathbf{G} \leftarrow \exp(\mathbf{C}/\beta)$
  for $t = 1, 2, 3 \ldots N$ **do**
    $\mathbf{Q} \leftarrow \boldsymbol{\gamma}^{(t)} \odot \mathbf{G}$ `// Hadamard product`
    for $k = 1, \ldots K$ **do** `// Set` $K = 1$ `in practice`
      $\boldsymbol{a} \leftarrow \frac{\mu_s}{\mathbf{Q}\boldsymbol{b}}$, $\boldsymbol{b} \leftarrow \frac{\mu_t}{\mathbf{Q}^T\boldsymbol{a}}$
    end
    $\boldsymbol{\gamma}^{(t+1)} = \mathrm{diag}(\boldsymbol{a})\mathbf{Q}\mathrm{diag}(\boldsymbol{b})$
  end
  $\widetilde{\boldsymbol{\gamma}} \leftarrow \boldsymbol{\gamma}^{(N+1)}$
return $\widetilde{\boldsymbol{\gamma}}$

OT solution where the step size is controlled by β. The implementation details for IPOT are in Algorithm 2. Using the approximated OT matrix, we define the loss of the distillation as the total transportation cost:

$$loss := \left\langle \widetilde{\boldsymbol{\gamma}}, \mathbf{C} \right\rangle_F \tag{4.5}$$

During training, we hold the teacher query encoder constant by removing its parameters from the computational graph and use the loss to update the student query encoder. Given each pair of $\hat{q}$ and q, minimizing the loss will lead the model to reduce the cosine distance between tokens according to the transportation plan. And because E_{M_q} is fixed, the essence of knowledge distillation is to push non-English token representations generated by E_{S_q} towards their corresponding English token representations generated by E_{M_q}. Moreover, though designed as the query encoder, the textual data of $\hat{q}$ and q used for distillation do not have to be the query from a CLIR dataset. A group of parallel sentences with a broad vocabulary coverage is adequate to train the student query encoder. Compared to the CLIR data, which

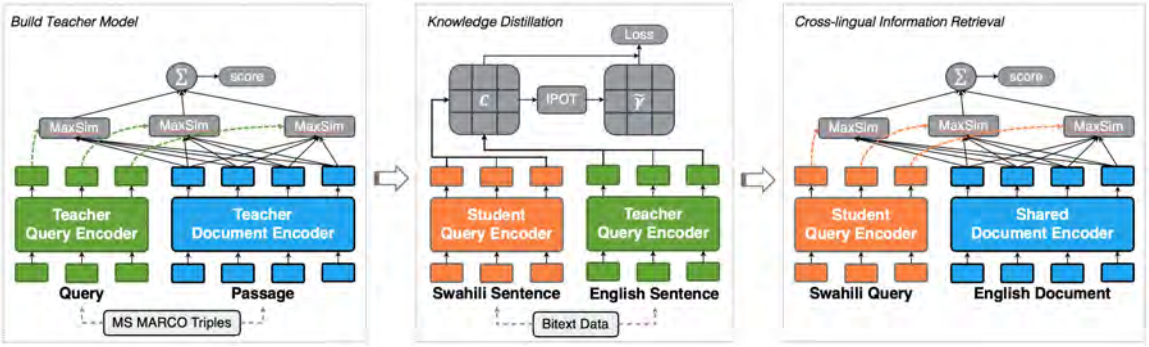


Figure 4.1: Model building pipeline for OPTICAL. This figure is based on CLIR task between Swahili query and English documents.

often require human relevance judgments, bitext data are easier to acquire, especially for low-resource languages.

4.1.3 Cross-lingual Query Document Matching

In this section, we focus on a CLIR task of searching English documents using a non-English query to introduce the OPTICAL framework. The document encoder in the student model can be directly copied from the teacher model ($E_{S_d} \leftarrow E_{M_d}$). At test time, the matching score of $\hat{q}$ and d is calculated based on equation (4.1) using E_{S_q} and E_{S_d}. A complete overview of the model building pipeline is shown in Figure 4.1. In fact, OPTICAL can be extended to different language settings in the CLIR task. For example, suppose the task requires searching non-English documents using an English query. In this case, the student model can reuse the teacher's query encoder and train a non-English document encoder using knowledge distillation. More generally, if the query is in X and the collection is in Y, where X and Y are both non-English languages, we can build the student model by training two knowledge distillations: X to English for query encoder and Y to English for document encoder.

4.2 Experimental Setup

4.2.1 Dataset

CLIR Settings. We focus on searching English collections with queries in low-resource languages (Cieri et al., 2016). We consider four low-resource languages in our experiments: Swahili, Somali, Tagalog, and Marathi. According to linguistic classification[1], they belong to four different language families: Niger-Congo (Swahili), Afro-Asiatic (Somali), Austronesian (Tagalog), and Indo-European (Marathi). To fully evaluate the performance of the proposed method, we also include three medium to high-resource languages: Finnish, German, and French.

Evaluation Data. We create a unified evaluation dataset for all language pairs considered in our experiments. The data are from the Cross-Language Evaluation Forum (CLEF) 2000-2003 campaign for bilingual ad-hoc retrieval tracks (Braschler, 2002b). The query is a concatenation of the title and description fields of the topic files. In total, there are 151 queries from the CLEF C001 – C200 topic (queries with no relevant judgments are removed). The collection of English documents is the Los Angeles Times corpus, composed of 113K news articles. For Finnish, German, and French, their queries are provided by CLEF campaign. For low-resource languages, Bonab et al. (2019) provided Somali and Swahili translations of English queries. And we hired bilingual human experts from Gengo[2] service to translate English queries into Tagalog and Marathi.

Retrieval Training Data. To guarantee a consistent performance of the teacher model on the monolingual retrieval task, we randomly sample a subset of 7 million triples from the MS MARCO passage ranking dataset for the training of the teacher model. The baselines, which involve synthesizing the CLIR dataset with different methods, all use the same subset of triples.

---

[1] `https://en.wikipedia.org/wiki/List_of_language_families` as of 05/01/2023.

[2] `https://gengo.com`

Bitext Data. To support the cross-lingual knowledge distillation, we use the parallel sentences from the CCAligned dataset (El-Kishky et al., 2020a). In our exploration of re-ranking comparison (Section 4.3.2), the distillation is trained based on a random sample of up to 2 million parallel sentences for each language pair. Note that for Somali and Marathi, the total number of parallel sentences in CCAligned is less than 2 million. Thus, we use all the parallel sentences available for these two languages (360K for Somali and 750K for Marathi). Moreover, there are other parallel corpora for the languages studied in our experiments that could help us to create larger training data. We use only CCAligned to ensure the consistency of the data quality in our experiments.

4.2.2 Implementation Details

We initialize the ColBERT query and document encoder components in both teacher and student models using pre-trained mBERT. We set the max length of all query encoders at $L = 32$ and truncate the document at 180 tokens. There are two model training tasks in our experiments. One is the **retrieval** training task. This is the main task of training the teacher model and the other neural baselines. Given a query, relevant passage, and non-relevant passage triplet, the models are trained using pairwise cross-entropy loss with a learning rate of 3×10^{-6} and a batch size of 64 for 200K iterations. The other training task is **knowledge distillation**. In this task we train the student model using bitext data. We set the step size to $\beta = 0.5$ and number of iterations to $N = 100$ for the IPOT algorithm. We use the cost of the optimal transportation as the loss and build a batch size of 256 with gradient accumulation. The student model is trained with a learning rate of 5×10^{-5} for 3 epochs of the available bitext data. Regarding the NMT models used in some of our

baseline methods, we use the off-the-shelf OPUS-MT (Tiedemann and Thottingal, 2020) from the Helsinki NLP group[3].

Evaluation. While we train models on passages for the retrieval task, our goal is to rank documents whose length is usually longer than 180 tokens. We split large documents into overlapping passages of fixed length with a stride of 90 tokens and compute the score for each query passage pair. Finally, we select a document's maximum passage score as its document score (Nair et al., 2022). Again, for CLEF dataset, we report the mean Average Precision (mAP) and the precision of the top 10 (P@10) ranked documents. We determine statistical significance using the two-tailed paired t-test with p-value less than 0.05 (i.e., 95% confidence level).

4.2.3 Compared Methods

First-Stage Retrieval. We employ a two-stage retrieval approach for addressing the CLIR problem, where first we obtain an initial set of candidate documents using a lexical matching retrieval technique (i.e., Okapi BM25) and then re-rank the initial set of candidate documents using a neural re-ranker. We select Recall@100 as our primary evaluation metric for the first-stage retrieval to collect the most relevant documents. We investigate different strategies for our initial ranking stage that we elaborate in the following:

- **SMT+BM25**: We translate the query based on a Statistical Machine Translation (SMT) method. More specifically, we first build a translation table from the CCAligned for each language pair using the GIZA++ toolkit. Then we select the top 10 translations from the translation table for each query term and apply Galago's weighted *#combine* operator to form a translated query. Finally, we run BM25 with default parameters to retrieve documents.

[3] `https://huggingface.co/Helsinki-NLP`

- **NMT+BM25**: Neural machine translation models based on the encoder-decoder architecture are empirically better than SMT in terms of translation quality (Yao et al., 2020b). Thus, we build this baseline by first translating the query into English using an NMT model. Then we run BM25 with default parameters to retrieve documents.

- **Human+BM25**: For a comprehensive comparison, we also provide an empirical upper bound on the initial ranking stage. We use CLEF English queries as the human translations and apply BM25 as the retrieval technique.

Neural Re-ranking. In the second retrieval stage, to provide the best initial rank list for neural re-ranking, we select the rank list from either SMT+BM25 or NMT+BM25 based on their Recall for each CLIR language pair. Including the proposed method, all the neural models rerank the top 100 documents of the initial rank list. We compare OPTICAL with the following methods:

- **mColBERT**: Because the encoders in the teacher model are based on a multilingual pre-trained language model, we can directly run the teacher model on the CLIR evaluation data in a zero-shot setting.

- **Code-Switch**: There are data augmentation methods that can help the training of cross-lingual tasks. Qin et al. (2021) proposed a code-switching framework to transform the monolingual training data into data in mixed languages. We apply the code-switch method to the queries in MS MARCO triples. More specifically, we randomly switch 50% of the English query words into their translations in the target language according to the Panlex dictionary.

- **Translate-Train**: Bonifacio et al. (2021) built a multilingual passage ranking dataset, mMARCO, by translating MS MARCO into target languages using the correspondent OPUS-MT models (Tiedemann and Thottingal, 2020). Nair

Table 4.1: Size of language data resource and OPUS-MT model performance.

(a) OPUS-MT model performance on low resource languages.

NMT Models	Swahili (SW)		Somali (SO)		Tagalog (TL)		Marathi (MR)	
Train./eval. data available	9M/386		0.8M/4		8M/2,500		5M/10,369	
Translation direction	EN-SW	SW-EN	EN-SO	SO-EN	EN-TL	TL-EN	EN-MR	MR-EN
BLEU scores	26.0	31.3	16.0	23.6	26.5	35.0	18.2	29.8

(b) OPUS-MT model performance on medium or high resource languages.

NMT Models	Finnish (FI)		German (DE)		French (FR)	
Train./eval. data available	45M/10,690		86M/17,565		180M/12,681	
Translation direction	EN-FI	FI-EN	DE-EN	EN-DE	EN-FR	FR-EN
BLEU scores	40.4	50.9	47.3	55.4	50.5	57.5

et al. (2022) showed that retrieval models trained using this synthetically generated CLIR dataset could outperform BM25 with query translation and the zero-shot neural approach in high-resource languages. We adopt this method as another baseline.

- **Translate-Test**: Like NMT+BM25, we can first let the NMT model translate the evaluation query into English and then perform English-to-English query document matching using a well-trained English retrieval model (i.e., the teacher model).

- **Human+ColBERT**: We also provide an empirical upper bound on the re-ranking stage. We use human translations of the evaluation query and apply the teacher model to re-rank the top 100 documents from the rank lists generated by the Human+BM25.

4.3 Results

4.3.1 First-Stage Retrieval Comparison

Table 4.2 shows the results of our first-stage retrieval methods. We can see that the NMT+BM25 approach outperforms SMT+BM25 in Recall@100 for all languages

Table 4.2: First-stage retrieval comparison. For recall columns, the highest value is marked with bold text. Note that the first row is an upper-bound reference.

(a) First-stage retrieval comparison on low resource languages.

First-Stage Retrieval	Swahili		Somali		Tagalog		Marathi	
	mAP	Recall	mAP	Recall	mAP	Recall	mAP	Recall
Human+BM25	0.4569	0.7621	0.4569	0.7621	0.4569	0.7621	0.4569	0.7621
SMT+BM25	0.2184	0.4359	0.1948	**0.4254**	0.1636	0.6195	0.1059	0.3289
NMT+BM25	0.2135	**0.4934**	0.1466	0.3670	0.3501	**0.6799**	0.1795	**0.4277**

(b) First-stage retrieval comparison on medium or high resource languages.

First-Stage Retrieval	Finnish		German		French	
	mAP	Recall	mAP	Recall	mAP	Recall
Human+BM25	0.4569	0.7621	0.4569	0.7621	0.4569	0.7621
SMT+BM25	0.3052	0.6049	0.3906	0.6946	0.4037	0.7541
NMT+BM25	0.3753	**0.7248**	0.4087	**0.7420**	0.4315	**0.7585**

except Somali. Referring to the evaluation in Table 4.1, the failure of NMT+BM25 on Somali is mainly due to poor translation quality. Moreover, human translation performs better than machine translation, and the margin is larger in low-resource (Table 4.1a) than high-resource (Table 4.1b) languages, indicating higher difficulty in building machine translation systems in low-resource languages. Based on recall, we choose the ranked lists from NMT + BM25 for the reranking step for all languages except Somali, for which we use SMT+BM25 as the initial retrieval method.

4.3.2 Re-ranking Comparison

Table 4.3 lists the evaluation results of both the first-stage retrieval methods and neural re-ranking models. For the zero-shot setting, we can see that mColBERT can improve the initial ranking on high-resource languages while it fails on low-resource languages. Similar to other downstream tasks (Wang et al., 2020; Wu and Dredze, 2020), the CLIR model based on a multilingual pre-trained language model also inherits the language bias in the pre-training step, causing the performance gap between low and high resource languages in document ranking. Using dictionary knowledge

Table 4.3: A comparison of model performance. ▷ are reported as the upper bound reference. The highest value is marked with bold text. Statistically significant improvements are marked by † (over Translate-Train) and ‡ (over Translate-Test).

(a) Model performance on low resource languages.

Retrieval Methods	Swahili		Somali		Tagalog		Marathi	
	mAP	P@10	mAP	P@10	mAP	P@10	mAP	P@10
▷ Human+BM25	0.4569	0.3940	0.4569	0.3940	0.4569	0.3940	0.4569	0.3940
SMT+BM25	0.2184	0.2152	0.1948	0.1865	0.1636	0.0934	0.1059	0.0984
NMT+BM25	0.2135	0.2113	0.1466	0.1380	0.3501	0.3179	0.1795	0.1795
▷ Human+ColBERT	0.5019	0.4344	0.5019	0.4344	0.5019	0.4344	0.5019	0.4344
mColBERT	0.1953	0.1795	0.1355	0.1212	0.3414	0.2960	0.1448	0.1556
Code-Switch	0.2420	0.2258	0.1845	0.1682	0.3542	0.2934	0.1573	0.1662
Translate-Train	0.2234	0.2185	0.1707	0.1649	0.3692	0.3252	0.1619	0.1722
Translate-Test	0.2643	0.2530	0.2126	0.2086	0.3827	0.3339	0.2141	0.2258
OPTICAL	**0.3129**†‡	**0.2901**†‡	**0.2477**†‡	**0.2365**†‡	**0.4188**†‡	**0.3623**†‡	**0.2414**†‡	**0.2384**†

(b) Model performance on medium or high resource languages.

Retrieval Methods	Finnish		German		French	
	mAP	P@10	mAP	P@10	mAP	P@10
▷ Human+BM25	0.4569	0.3940	0.4569	0.3940	0.4569	0.3940
SMT+BM25	0.3052	0.2821	0.3906	0.3437	0.4037	0.3772
NMT+BM25	0.3753	0.3583	0.4087	0.3580	0.4315	0.3881
▷ Human+ColBERT	0.5019	0.4344	0.5019	0.4344	0.5019	0.4344
mColBERT	0.3791	0.3272	0.4509	0.3807	0.4512	0.3868
Code-Switch	0.3831	0.3404	0.4553	0.3827	0.4589	0.3993
Translate-Train	0.4043	0.3576	0.4713	0.3967	0.4666	0.4020
Translate-Test	**0.4418**	**0.4024**	0.4811	**0.4080**	**0.4984**	**0.4318**
OPTICAL	0.4228	0.3874†	**0.4832**	0.4067	0.4764	0.4119

for cross-lingual data augmentation, the Code-Switch method performs better than mColBERT. However, the word-level translation knowledge used in Code-Switch does not consider the context of the switched terms, which could cause the semantics of the code-switched data to diverge from the original one. Comparing Code-Switch to Translate-Train, we can see that Translate-Train outperforms Code-Switch in high-resource languages. With the support of the NMT model, the Translate-Train method can generate better query translations in high-resource languages, which leads to a higher quality of CLIR training triples than the Code-Switch method. However, in low-resource languages Swahili and Somali, Translate-Train cannot consistently

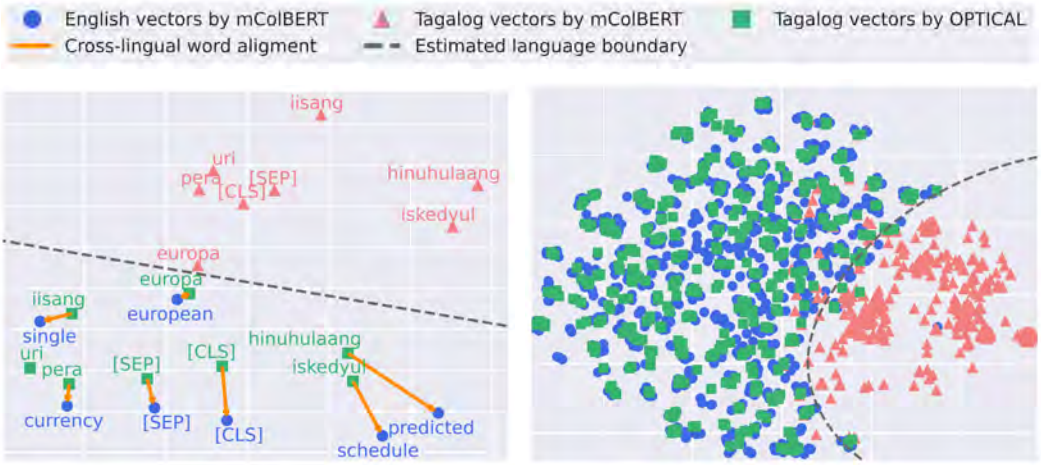
outperform Code-Switch. This is because the NMT model does not have enough training resources, and the generated query translations are of lower quality. Instead of building a CLIR dataset for model training, the Translate-Test method translates the query to English using an NMT model and then ranks the document based on a monolingual neural retrieval model. With the help of two neural models at test time (translation and document ranking), this two-step approach becomes the strongest baseline in our experiment.

Finally, we can see the effectiveness of OPTICAL in low-resource languages. Our method consistently and significantly improves the first-stage retrieval results. In low-resource languages, OPTICAL substantially outperforms all baselines, including the Translate-Test method. In high-resource languages, OPTICAL also achieves solid performance. It outperforms the Translate-Train method in all three languages. And it is a surprise to us that OPTICAL exceeds the Translate-Test on German in terms of mAP. Moreover, the results of OPTICAL in Table 4.3 are only based on a maximum of 2M parallel sentences (there are only 360K for Somali and 750K for Marathi). No cross-lingual relevance judgment is used in the distillation step, making OPTICAL data feasible and easy to build. At the same time, we can see that using human translation with a neural ranking model (Human+ColBERT) still leads the CLIR setting with the same model architecture by a large margin in low-resource languages.

4.3.3 Analysis of Knowledge Distillation

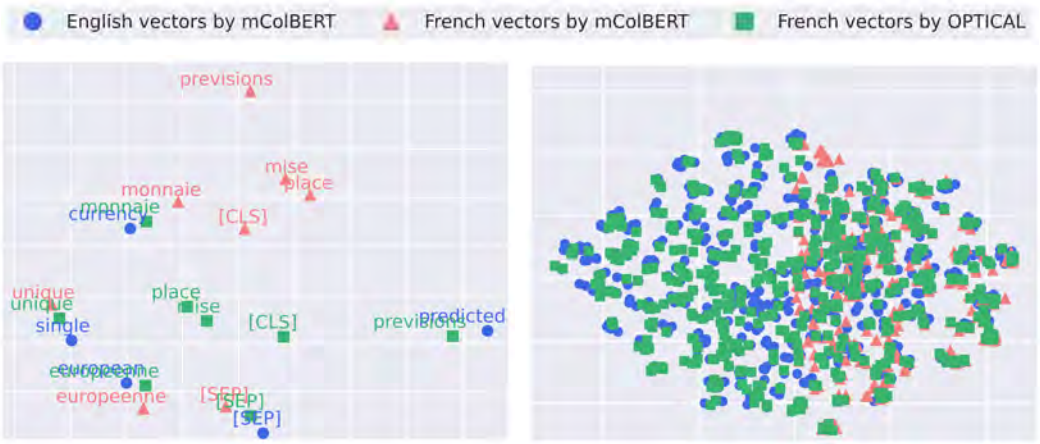
To study how the student model behaves after OPTICAL distillation, we compare the token representations of the same query generated by different models.

Student query encoder in low-resource language. We consider three types of token representations. First, we encode the English query using the mColBERT query encoder. Note that the mColBERT query encoder is the same as the OPTICAL teacher encoder, which provides the knowledge to the student query encoder



(a) Example query comparison.

(b) Overview of all test queries.



(c) Example query comparison.

(d) Overview of all test queries.

Figure 4.2: t-SNE visualization of query tokens. This group of figures compares Tagalog (upper) and French (bottom) with the same example query (left) and set of queries (right).

during the distillation. Then we encode the same corresponding Tagalog query using both mColBERT and OPTICAL student query encoders. Finally, we use t-SNE (Van der Maaten and Hinton, 2008) to project these high-dimensional vectors to two-dimensional space. Figure 4.2a visualizes an example query. In English, the query is *"What is the schedule predicted for the European single currency?"* The parallel Tagalog query is *"Ano ang hinuhulaang iskedyul para sa iisang uri ng pera sa Europa?"* Figure 4.2b provides an overview of all test queries in both English and

Tagalog. We can see that when used in a zero-shot setting, the English and Tagalog token representations generated by the teacher model have a clear language boundary. Although starting from a multilingual pre-trained language model, mColBERT is only trained on MS MARCO English data, so only English tokens have knowledge of query document matching. Therefore, we observe (in Table 4.3a) a large retrieval performance gap on the same model between the English query (i.e., Translate-Test) and the query in the low-resource languages (i.e., mColBERT). On the other hand, Tagalog token representations generated by the student encoder are much closer to English token representations. More importantly, by reducing the transportation cost of the parallel sentences, OPTICAL can push Tagalog word vectors toward their corresponding English vectors. Eventually, Tagalog words that are close to their English translations can also obtain retrieval knowledge because the English token representations are generated by the teacher model. This matches the design purpose of OPTICAL.

Student query encoder in high-resource language. We repeat the same analysis but for French queries. Figure 4.2c shows the t-SNE visualization of the same query in French (*Quelles sont les prévisions pour la mise en place de la monnaie unique européenne?*). Figure 4.2d provides the overview of all test queries in both English and French. Different from low-resource languages, we can see that English and French words are already mixed in the representation space, so there is no clear language boundary. Tokens in French are already close to their translations in English. This explains why the effect of knowledge distillation from OPTICAL is more significant in Tagalog than in French.

4.3.4 Effect of Bitext Data Size

OPTICAL results in Table 4.3 are based on a maximum of 2M bitext data used for training distillation. In this experiment, we study the effect of bitext data size on

Figure 4.3: Performance with respect to bitext data size.

OPTICAL. We select two low-resource languages with a relatively larger collection of parallel sentences in CCAligned: Swahili (2M) and Tagalog (6.6M). Then we train OPTICAL using different sizes of the bitext data: 100K, 500K, 1M, 2M, and for Tagalog, we also experiment on 4M bitext data. Figure 4.3 shows the mAP performance with respect to the size of bitext data. As expected, more bitext data with larger vocabulary size and broader semantic coverage lead to better reranking performance. For Swahili, OPTICAL exceeds the Translate-Test method with a set of only 100K parallel sentences. For Tagalog, starting from 500K, OPTICAL performs better than Translate-Test. This demonstrates that OPTICAL is data-efficient.

4.3.5 Reduce High-resource to Low-resource

We hypothesize that the strong performance of the Translate-Test method on high-resource languages is mainly because of the excellent translation quality from the NMT models. Yet a large amount of training data is a prerequisite for the success of NMT. In this experiment, we turn high-resource to low-resource language by limiting the training data size of the NMT models. For French and German, the OPUS corpora have more than 800M parallel data in total, and the CCAligned dataset alone has 15M pairs for each language. We follow the same architecture of the

Table 4.4: mAP comparison of reducing high-resource to low-resource.

Retrieval Methods	Limited size of NMT model training			
	French		German	
	5M	10M	5M	10M
Translate-Test	0.3820 (-19.8%)	0.4525 (-5.0%)	0.3971 (-17.8%)	0.4667 (-3.4%)
OPTICAL (2M)	**0.4764**		**0.4832**	

OPUS-MT model but only use subsets with sizes of 5M and 10M pairs, respectively, from CCAligned for training. We compare Translate-Test with the suboptimal NMT models and OPTICAL in Table 4.4. The drops on mAP of Translate-Test indicate the performance of the Translate-Test heavily relies on the NMT model which is data-hungry. Moreover, the knowledge distillation step in OPTICAL and the training of the NMT model use the same type of data, i.e., the bitext data. Therefore, for learning the translation knowledge in the CLIR task, OPTICAL is more data efficient than the NMT model.

4.4 Summary

This chapter addresses the data scarcity issue for training CLIR models. We present OPTICAL, an optimal transport knowledge distillation framework for CLIR tasks involving low-resource languages. OPTICAL builds CLIR models by separating the retrieval knowledge from the translation knowledge. First, the teacher model learns the retrieval knowledge in a monolingual setting. Then we design a knowledge distillation loss based on the optimal transport costs to transfer the retrieval knowledge to the student model in a cross-lingual setting. We achieve the cross-lingual transfer only based on bitext data which greatly reduces the data required for building CLIR models. Our comprehensive experimental results show that OPTICAL significantly outperforms other baselines in low-resource languages, including

the NMT models. Further analysis demonstrates the effectiveness of the knowledge distillation step in OPTICAL.

In the next chapter, we expand the idea of cross-lingual knowledge transfer from CLIR to MLIR. We will focus on searching a multilingual collection using English queries and explore a model architecture that can unify the representation of documents from different languages into a monolingual (e.g., English) embedding space.

CHAPTER 5

LANGUAGE PROMPT FOR MULTILINGUAL KNOWLEDGE TRANSFER

In this chapter, we expand the scope of CLIR to MLIR, allowing both query and document to be expressed in any language (or within a group of languages). Considering the language setting in CLIR as a one-to-one scenario, we refer to the language setting in MLIR as the many-to-many scenario. The extension of language setting creates two major challenges: (i) From a language perspective, the translation component in the MLIR models needs cross-lingual knowledge for multiple language pairs. (ii) From a retrieval perspective, the ranking component in the MLIR models needs to perform consistently across languages. Motivated by many real-world applications, such as web search, where the retrieval collection includes documents from multiple languages (Chaware and Rao, 2009), we first focus on one-to-many setting of MLIR, where the query is in English, and the collection is a mixture of languages.

The *data scarcity* and *language bias* issues in CLIR become more challenging in the context of MLIR. First, to learn query-document matching knowledge on multiple language pairs effectively, a model built using multilingual relevance judgment requires access to *retrieval* training data that covers the languages present in the target collection. However, the scarcity of training data between two languages in CLIR is exacerbated with multiple language pairs in MLIR (Lawrie et al., 2022). Therefore, it is preventing training to achieve broad language coverage. Second, due to the unbalanced pre-training data in different languages, language bias causes performance to vary from one CLIR task to another. However, Xu et al. (2001) found that in MLIR, the distribution of relevant documents to a given query often differs

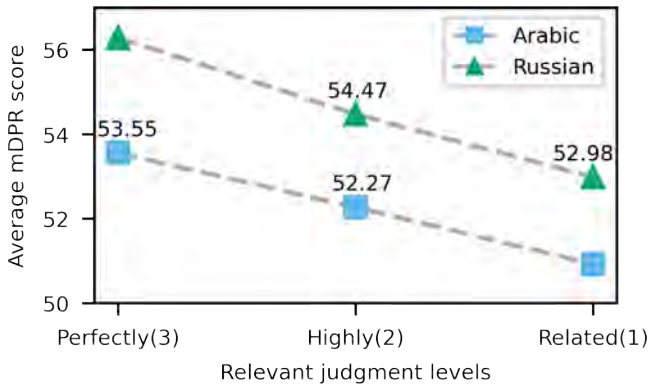


Figure 5.1: Average score given to parallel documents in Arabic and Russian by mDPR (Zhang et al., 2021). Documents are translated by mMARCO (Bonifacio et al., 2021).

in different languages – which highlights the challenges in designing effective MLIR models that also perform fairly across languages. Therefore, language bias leads to inconsistent ranking results in MLIR tasks, even among high-resource languages. To demonstrate this case, we pair the test queries from TREC 2020 Deep Learning Track (Craswell et al., 2020) with their relevant passages translated into Arabic and Russian by mMARCO (Bonifacio et al., 2021). Then for each language, we score query-document pairs using mDPR (Zhang et al., 2021). Figure 5.1 illustrates the difference in ranking the same set of relevant documents in these two languages. We observe that mDPR scores Russian documents higher than their Arabic version. We argue that such inconsistency in MLIR would lead to sub-optimal ranking results, e.g., highly relevant documents in Arabic have lower scores than slightly relevant documents in Russian.

In the previous chapter, we learned that it is possible to transfer a monolingual retrieval model to a target language through cross-lingual knowledge distillation. To address additional challenges in MLIR setting, we could transfer a monolingual (e.g., English) retrieval model to multiple languages using multi-task knowledge distillation. Each task is a cross-lingual knowledge transfer from English to one language

in the collection. Following this idea, we present KD-SPD,[1] a multilingual dense retrieval model based on knowledge distillation (KD) and soft prompt decoder (SPD) for the MLIR task. KD-SPD does not require any multilingual relevance labels for training, thus automatically solving the data scarcity issue in low-resource languages. Our approach solely requires monolingual retrieval training data in English, which we obtain from MS MARCO (Nguyen et al., 2016), and a large collection of parallel and comparable documents for cross-lingual knowledge transfer. We first adopt a dense passage retrieval (DPR) model built for English as the teacher model. Unlike token-level representations in the ColBERT architecture, the representation of the DPR model is at the document level. The ranking score is computed as the dot-product of representations between the query and document. Since the teacher has the ability to do query-document matching, we freeze its document encoder and then minimize the distance between the representations generated by the teacher for any English document and the representations learned by KD-SPD for its parallel or comparable version in other languages. Therefore, our approach implicitly "translates" the representation of documents in different languages into the same language embedding space. Moreover, we hypothesize that although different languages possess unique properties such as distinct grammar or vocabulary, they also have common traits for expressing similar meanings. To capture these unique and shared features, KD-SPD uses a decomposable soft prompt, which is derived as the product of a shared matrix and a low-rank language-specific matrix for each language. Through joint training across multiple languages, we observe that the learned prompts are capable of reducing language bias and possess the transferable capacity to generalize to unseen languages.

[1]KD refers to the model training framework, and SPD refers to the model architecture.

The work described in this chapter, namely Soft Prompt Decoding for Multilingual Dense Retrieval, was published in SIGIR 2023 (Huang et al., 2023b). I was the lead author who designed the model architecture and conducted the experiments.

5.1 Design Overview

Given a query q in language X and a target collection $\mathcal{D}_{\boldsymbol{Y}}$ which contains documents in language set $\boldsymbol{Y} = \{Y_1, Y_2, \dots Y_K\}$, suppose d_{ki}—the i^{th} document in language Y_k—has the ground truth relevance label $Rel(q, d_{ki})$, then the aim is to design an MLIR model f that retrieves a list of documents from $\mathcal{D}_{\boldsymbol{Y}}$ such that

$$f(q, d_{ki}) \geq f(q, d_{lj}), \quad \forall\ Rel(q, d_{ki}) \geq Rel(q, d_{lj}) \tag{5.1}$$

where $f(\cdot, \cdot)$ indicates the ranking score calculated by the model. To build model f, we first assume there exists an oracle model g for the retrieval task in language X. Thus, given q and monolingual collection $\mathcal{D}_X$, g satisfies:

$$g(q, d_{xi}) \geq g(q, d_{xj}), \quad \forall\ Rel(q, d_{xi}) \geq Rel(q, d_{xj}) \tag{5.2}$$

We can achieve (5.1) with model f' if for any d_* in $\boldsymbol{Y}$ and its translation d_x in X, the model matches the oracle:

$$f'(q, d_*) = g(q, d_x)$$

Suppose both f' and g follow the architecture of dense retrieval, the ranking score calculation is the dot-product of the query and document embeddings, thus:

$$f'_E(q) f'_D(d_*)^\top = g_E(q) g_D(d_x)^\top$$

where f'_E and g_E are query encoders; f'_D and g_D are document encoders for f' and g respectively. We then reuse g_E as the query encoder of f'. With $f'_E = g_E$, we have:

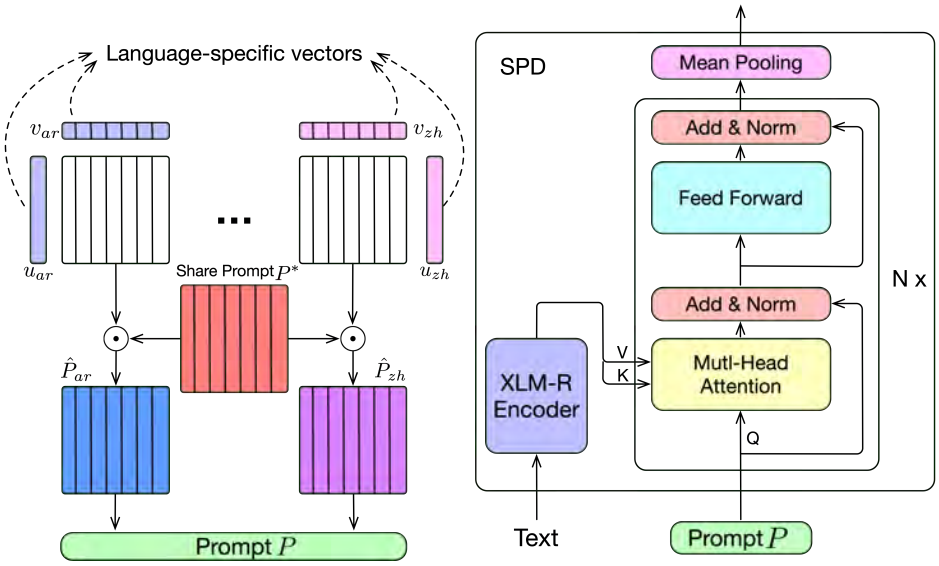
$$g_E(q)\big(f'_D(d_*) - g_D(d_x)\big)^\top = 0 \tag{5.3}$$

It is safe to assume $g_E(q)$ is a nonzero vector. Therefore the goal of finding f' is equivalent to reducing the embedding distance between parallel documents. In our method, we retrain g_D as the teacher model by removing its parameters from the computational graph and train f'_D as the student model.

Note that in practice, the oracle model g does not exist. We can use an off-the-shelf English-to-English (monolingual) dense retrieval model as a substitute for g. Because g_D is fixed, the essence of knowledge distillation training is to push multilingual document representations generated by f'_D toward their corresponding English document representations generated by g_D. Moreover, equation (5.3) suggests that the training of f'_D does not rely on either query q or ground truth relevant judgment. A group of parallel or comparable sentences from English to any other language involved in the collection is adequate to train f'_D. Parallel or comparable sentences between two languages are often referred as bitext data. Unlike multilingual retrieval data, which often require relevance labels, bitext data are easier to acquire, especially for low-resource languages (Heffernan et al., 2022).

5.2 Soft Prompt Decoder

We focus on the design of the document encoder of the student model, f'_D, which handles multilingual documents. In general, the function of f'_D is similar to a neural machine translation model. The difference is that f'_D translates the input text into an embedding in the target language rather than into natural language text. Thus, we build f'_D based on the encoder-decoder architecture. For the encoder component



(a) Soft prompt decomposition. (b) Multilingual document encoder.

Figure 5.2: SPD model architecture.

of f_D, we exploit multilingual pre-trained language models (i.e., mBERT or XLM-R). The token representation generated by the encoder is then forwarded to the decoder component. However, unlike the decoder with an auto-regressive generation process, we present a soft prompt-based decoder (SPD) architecture.

5.2.1 Soft Prompt Matrix

We consider f'_D a multitask model where translating (mapping) each language in the multilingual collection into the target language space is viewed as a single task. Using the language name as the task identifier, a prompt $\mathbf{P}_k \in \mathbb{R}^{l \times d}$ for language Y_k with the same dimension as the token embedding d and vector length as l is used as input to the decoder. Thus, the prompt matrix serves as the language-based decoding initialization vector. Inspired by prompt decomposition from multitask prompt tuning (Wang et al., 2023), we decompose $\mathbf{P}_k$ into two parts, as shown in Figure 5.2a: language-specific low-rank vectors $\mathbf{u}_k \in \mathbb{R}^l$ and $\mathbf{v}_k \in \mathbb{R}^d$ for language Y_k (in the figure, $Y_k = ar$ or $Y_k = zh$), and a shared prompt $\mathbf{P}^* \in \mathbb{R}^{l \times d}$ across all

languages. The language-specific prompt can be parameterized as $\mathbf{W}_k = \mathbf{u}_k \cdot \mathbf{v}_k^\top$, which has the same dimension as the shared prompt $\mathbf{P}^*$. The final prompt $\hat{\mathbf{P}}_k$ for language Y_k is then formulated as follows.

$$\hat{\mathbf{P}}_k = \mathbf{P}^* \odot \mathbf{W}_k = \mathbf{P}^* \odot (\mathbf{u}_k \cdot \mathbf{v}_k^\top) \tag{5.4}$$

where $\odot$ denotes the element-wise product between two matrices. The shared prompt enables efficient knowledge sharing across all source languages and commonalities across translation tasks. Meanwhile, the language-specific vectors allow each translation task to maintain its parameters to encode language-specific knowledge. Additionally, prior studies on multitask prompt learning also showed that soft prompt learned from multitask data could be efficiently transferred to a new task (Su et al., 2022; Vu et al., 2021). In section 5.5.4, we show that with a shared prompt, the SPD has a better zero-shot transfer ability toward new languages.

5.2.2 Cross-attention Decoder

The decoder network follows a cross-attention-based multi-layer transformer architecture. Each layer has two sub-layers. The first is a multi-head query-key-value (QKV) cross-attention module, and the second is a position-wise fully connected feed-forward network. We employ residual connection and layer norm around each of the sub-layers.

Let $\mathbf{T}_{d_k} \in \mathbb{R}^{|d_k| \times d}$ denote the token representations generated by the encoder component for document d_k in language Y_k, where $|d_k|$ is the number of tokens in d_k. The first decoder layer applies the cross-attention module between $\mathbf{T}_{d_k}$ and prompt matrix $\hat{\mathbf{P}}_k$. On the mth head, the attention mechanism is defined as follows:

$$\text{Attention}_m = \text{Softmax}\Big(\frac{W_m^q \hat{\mathbf{P}}_k \cdot W_m^k \mathbf{T}_{d_k}}{\sqrt{d/M}}\Big) W_m^v \mathbf{T}_{d_k}$$

where M is the number of heads and W_m^q, W_m^k and W_m^v are matrices with dimension $d/M \times d$. Thus, the prompt matrix has different attention weights over encoder token representations in each subspace projection (head). The output of the multi-head QKV cross-attention module is the concatenation of M heads with linear projection:

$$\text{CrossAttention}(\hat{\mathbf{P}}_k, \mathbf{T}_{d_k}) = W^o[\text{Attention}_1, \ldots, \text{Attention}_M]$$

We further define the output of the attention-based sub-layer with the residual connection and layer norm:

$$\mathbf{h}_{d_k} = \text{LN}\big(\hat{\mathbf{P}}_k + \text{CrossAttention}(\hat{\mathbf{P}}_k, \mathbf{T}_{d_k})\big)$$

where $\text{LN}(\cdot)$ denotes the layer norm operation. Because $\hat{\mathbf{P}}_k$ is the query element in the cross-attention module, we use the prompt matrix to query the information from the encoder output and store it in a hidden representation $\mathbf{h}_{d_k}$ which has the same dimension as the prompt matrix. Next, we apply the second sub-layer and generate the output of the first decoder layer for d_k, $\mathbf{H}_{d_k}^1 \in \mathbb{R}^{l \times d}$:

$$\mathbf{H}_{d_k}^1 = \text{DecoderLayer}_1(\hat{\mathbf{P}}_k, \mathbf{T}_{d_k}) = \text{LN}(\mathbf{h}_{d_k} + \text{FFN}(\mathbf{h}_{d_k}))$$

where $\text{FFN}(\cdot)$ denotes the fully connected feed-forward network with a rectified activation function. Then we use the hidden representation from the previous layer (i.e. $\mathbf{H}_{d_k}^1$) to query the encoder output again in the next layer, that is:

$$\mathbf{H}_{d_k}^{n+1} = \text{DecoderLayer}_{n+1}(\mathbf{H}_{d_k}^n, \mathbf{T}_{d_k})$$

until reaching the maximum layer N designed for the decoder. Finally, we average $\mathbf{H}_{d_k}^N$ over the prompt vector dimension as the document embedding in the target

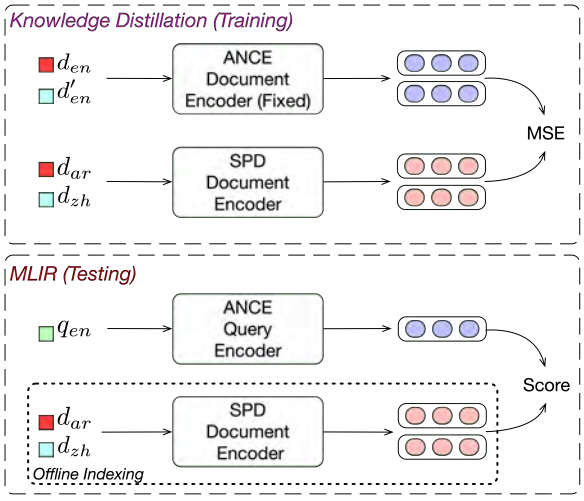


Figure 5.3: Model building pipeline for MLIR.

language space. A complete architecture of f_D is depicted in Figure 5.2b.

$$f'_D(d_k) = \text{MeanPool}(\mathbf{H}^N_{d_k})$$

5.3 Multilingual Dense Retrieval

Knowledge Distillation Training. Assume that d_{En} is the English version of d_k. From the property of the oracle g, we know that the document embedding of d_{En} generated by g_D contains rich knowledge for query-document matching in English. Equation (5.3) suggests that if we could let f'_D "behave" like g_D, namely, if for any d_k, the output of $f'_D(d_k)$ is close to the output of $g_D(d_{\text{En}})$, then the document embedding generated by f'_D can have a similar retrieval performance as g in the English domain. Therefore, we use the English document encoder g_D as the teacher model and our multilingual document encoder f'_D as the student. During training, we define the distillation loss as the mean square error (MSE) between two embeddings and sample B examples from each of K languages to form a batch.

$$loss := \frac{1}{KB} \sum_{k=1}^{K} \sum_{i=1}^{B} |f'_D(s_{ki}) - g_D(e_{ki})|^2 \tag{5.5}$$

where s_{ki} is a sentence in language Y_k and e_{ki} is its parallel (translation) in English.

Query-document matching. In this section, we discuss an MLIR task of searching multilingual collections using an English query to introduce the KD-SPD framework. The query encoder in the final retrieval model can be directly copied from the teacher model in the English domain. Specifically, at test time, the matching score of q and d_* is calculated based on the dot-product between g_E and f'_D:

$$f'(q, d_*) = g_E(q) f'_D(d_*)^\top$$

An overview of our MLIR model building pipeline is shown in Figure 5.3. In fact, we can also apply KD-SPD to other language settings in MLIR task. For example, suppose the task requires searching an English collection using queries in multiple languages. In this case, KD-SPD can be built as a query encoder, and the retrieval model can reuse the teacher's document encoder. More generally, if the MLIR task involves a query language set $\boldsymbol{X}$ and a collection language set $\boldsymbol{Y}$, we can consider English as a bridge to build KD-SPD via two knowledge distillations: $\boldsymbol{X}$ to English for query encoder and $\boldsymbol{Y}$ to English for document encoder.

5.4 Experimental Setup

5.4.1 Dataset

Evaluation Data. We focus on retrieval from multilingual collections with English queries. To comprehensively evaluate model performance on this MLIR task, we create three test sets with various combinations of collection size, relevance distribution, and language settings. Table 5.1 shows the statistics of our evaluation datasets.

- **CLEF.** The data is from the CLEF 2000-2003 campaign for bilingual ad-hoc retrieval tracks (Braschler, 2002b). We include documents in French, German, and Italian to build a multilingual collection. Among the CLEF C001 – C200 topics, we only consider a topic with human-annotated relevant documents in all three languages as a valid query, leading to 133 queries in total.

- **mTREC.** The query and relevance judgments are from the test split of the passage ranking task from the TREC 2020 Deep Learning Track Craswell et al. (2020). There are three relevance judgment levels marked by {3,2,1}, with 3 being most relevant. We build the multilingual collection from mMARCO (Bonifacio et al., 2021). We select translated passages in four languages: Arabic, Chinese, Russian, and Indonesian, to form a large-scale multilingual collection.

- **LAReQA.** LAReQA (Roy et al., 2020) is a benchmark for language-agnostic answer retrieval from a multilingual candidate pool. It is built based on two multilingual question-answering datasets: XQuAD Artetxe et al. (2019) and MLQA Lewis et al. (2019). The query is formed using the question, and the collection is formed by breaking contextual paragraphs into sentences. Each query (question) appears in 11 different languages[2] and has 11 parallel relevant sentences (answers). To match our MLIR setting, we evaluate English queries on a collection of sentences in 11 languages (including English).

Bitext Training Data. To support the multilingual knowledge distillation, we use the parallel sentences from the CCAligned dataset (El-Kishky et al., 2020a). To train one KD-SPD model covering all three evaluation datasets (15 languages[3]), we sample 4 million parallel sentences per language except English. For English, to be consistent with other languages, we sample another 4 million sentences and pair each

[2]Languages in LAReQA (ISO code): ar, de, el, en, es, hi, ru, th, tr, vi, zh

[3]List of training languages (ISO code): ar, de, el, en, es, fr, hi, id, it, pt, ru, th, tr, vi, zh

Table 5.1: Summary of MLIR evaluation datasets. Avg. $\#d^+/q$ denotes the average number of relevant documents per query.

Dataset Statistics	CLEF	mTREC	LAReQA
Number of Queries	133	54	1,190
Number of Docs	241K	35.2M	13,014
Languages in collection	3	4	11
Avg. $\#d^+/q$	13.5	66.8	1.0

sentence with itself. Thus, our training data comprises 60 million sentence pairs in 15 languages. We also append a language code to each sentence pair for SPD to identify the language of the input document.

Retrieval Fine-tuning Data. For a competitive baseline, we further fine-tune an mDPR baseline (see section 5.4.3) using cross-lingual triples from mMARCO. We sample 6 million cross-lingual triples per language to form a multilingual training set for languages in CLEF and mTREC. Because languages in LAReQA are not fully covered by mMARCO, we use mDPR on LAReQA without fine-tuning. Note that our KD-SPD model does not use this data.

5.4.2 Implementation Details

We initialize the encoder component of the SPD model using XLM-R model and the decoder component (including prompt matrices) using the Xavier initialization (Glorot and Bengio, 2010). We train the SPD as a student model using bitext data. To learn the retrieval knowledge in the English domain, we employ the document encoder of ANCE (Xiong et al., 2021) as the teacher. When testing, the query encoder of the final model is also a reuse of the query encoder of ANCE (except in section 5.5.3, where we investigate the impact of different teachers). For hyper-parameters, we set the length of the prompt token vector $l = 30$ and the number of SPD decoding layers $N = 6$. We truncate the input sequence length at 180 tokens and sample 4 examples per language to build a mini-batch. The model is trained with a learning rate of 2×10^{-5} for one epoch of all bitext data. For evaluation on the

CLEF dataset, where the document length is usually longer than 180 tokens, we split long documents into overlapping passages of fixed length with a stride of 90 tokens and compute the score for each query passage pair. Finally, we select a document's maximum passage score as its ranking score (Nair et al., 2022).

Evaluation. We examine the top 100 ranked documents and report comprehensive metrics. In addition to mean Average Precision (mAP) and precision of the top 10 (P@10) ranked documents, we also include normalized Discounted Cumulative Gain (nDCG@10), mean reciprocal rank (MRR), and recall (R@100). We determine statistical significance using the two-tailed paired t-test with p-value less than 0.05 (i.e., 95% confidence level).

5.4.3 Compared Methods

From a modeling perspective, we compare KD-SPD with both non-neural and neural approaches. From the system design perspective, we compare KD-SPD with end-to-end solutions and pipeline solutions via rank list merging. For non-neural baselines, we generally consider a three-step pipeline to address MLIR. First, we break the collection into subsets by language and translate the query to each subset language. Since the translated queries and subset collection are in the same language, we then use a lexical-based sparse retrieval technique (e.g., BM25) to obtain a ranked list for each language. Finally, we merge language-specific ranked lists into a final ranked list. We investigate different strategies of translation and ranked list merging that we elaborate below.

- **SMT.** We translate the query based on a Statistical Machine Translation (SMT) method. Specifically, we first build a translation table from the parallel corpus for each language pair using GIZA++. Then we select the top 10 translations from the translation table for each query term and apply Galago's *#combine*

operator to form a translated query. Finally, we run BM25 with default parameters to retrieve documents in the same language as the query translation.

- **GMT.** We translate the query into the collection languages using Google Translation[4]. Then, we run BM25 with default parameters to retrieve documents from each subset collection using the translated query. Note that because Google Translation is an evolving system, the results regarding GMT are not guaranteed to be reproducible.

- **+RR.** We merge multiple rank lists in the round-robin (RR) style, that is, iteratively extracting the top-ranked document from K languages in random order to be the next K of the final rank list.

- **+Score.** We merge multiple rank lists by ranking scores generated by the retrieval component. Scores within each rank list are min-max to range $[0, 1]$.

The non-neural baselines are the combination of translation with merging strategies: SMT+RR, SMT+Score, GMT+RR, and GMT+Score. As a dense retriever, we compare KD-SPD with other dense retrieval methods in the following:

- **mDPR.** Models that follow the dense passage retriever (DPR) paradigm have proven to be effective for many retrieval tasks. Zhang et al. (2021) extended DPR to non-English languages by changing the underlying pre-trained language model from BERT to multilingual BERT (mBERT). We adopt the checkpoint of mDPR trained on MS MARCO dataset (Nguyen et al., 2016). For CLEF and mTREC, which have fewer languages in the collections, we further fine-tune mDPR using the mMARCO dataset (Bonifacio et al., 2021). We apply mDPR to MLIR in two ways: First, we break the MLIR task into multiple CLIR tasks by language and use mDPR to retrieve documents from subset

[4] `https://translate.google.com/`

collections. Then we merge the rank lists from different CLIR tasks, named mDPR+RR and mDPR+Score, respectively. Second, we apply mDPR as an end-to-end solution for MLIR, in which we use it to directly index and search from the multilingual collection.

- **KD-Encoder.** There are methods that can transfer the knowledge from a model built for a monolingual task to a multilingual model, enabling it to address the same task in a multilingual setting. Reimers and Gurevych (2020) proposed a knowledge distillation method to create multilingual versions from the same monolingual models. We refer to this idea as the KD-Encoder and apply it to the MLIR task. To compare with our approach, we adopt the same teacher model and train KD-Encoder with the same bitext data.

5.5 Results

5.5.1 Retrieval Performance

Table 5.2 lists the evaluation results on the three MLIR datasets. Comparing non-neural approaches, we can see that methods based on GMT outperform those based on SMT. For document collections with mostly high-resource languages, the GMT-based methods can also achieve higher nDCG, precision, and MRR scores than end-to-end neural approaches (i.e., GMT+Score on CLEF). It highlights that translation quality is an important factor in MLIR.

Usually, for a pipeline approach, the error can accumulate for each step and lead to a sub-optimal result (Ferreira et al., 2019). In MLIR, without evaluating the content with respect to the query, merging rank lists only based on the score or rank within the sub-collection will cause errors from multiple languages to accumulate. However, comparing the pipeline with the end-to-end approach of mDPR, we can see that end-to-end mDPR does not show a consistent advantage over the pipeline

Table 5.2: A comparison of model performance. The highest value is marked with bold text. For KD-SPD, statistically significant improvements are marked by † (over mDPR) and ‡ (over KD-Encoder).

(a) Model performance on CLEF.

Methods	**CLEF**				
	mAP	nDCG@10	P@10	MRR	R@100
SMT+RR	0.1348	0.2540	0.2429	0.4017	0.3732
SMT+Score	0.1459	0.2737	0.2421	0.4679	0.3508
GMT+RR	0.1783	0.3732	0.3474	0.5793	0.4118
GMT+Score	0.1950	0.3806	0.3474	0.6140	0.4206
mDPR+RR	0.1823	0.3412	0.3165	0.5448	0.4330
mDPR+Score	0.1941	0.3433	0.3203	0.5364	0.4401
mDPR	0.2025	0.3466	0.3195	0.5367	0.4504
KD-Encoder	0.1973	0.3883	0.3594	0.5641	0.4315
KD-SPD	**0.2200**†‡	**0.4160**†‡	**0.3714**†‡	**0.6356**†‡	**0.4689**‡

(b) Model performance on mTREC.

Methods	**mTREC**				
	mAP	nDCG@10	P@10	MRR	R@100
SMT+RR	0.0242	0.0557	0.0630	0.1592	0.0778
SMT+Score	0.0187	0.0468	0.0648	0.1060	0.0661
GMT+RR	0.0653	0.1735	0.1870	0.3965	**0.1872**
GMT+Score	0.0522	0.1570	0.1685	0.3970	0.1691
mDPR+RR	0.0490	0.1358	0.1537	0.2913	0.1324
mDPR+Score	0.0492	0.1459	0.1574	0.3154	0.1300
mDPR	0.0549	0.1675	0.1870	0.3954	0.1291
KD-Encoder	0.0639	0.2208	0.2293	0.4556	0.1629
KD-SPD	**0.0748**†‡	**0.2414**†‡	**0.2556**†‡	**0.5067**†‡	0.1705†

(c) Model performance on LAReQA.

Methods	**LAReQA**				
	mAP	nDCG@10	P@10	MRR	R@100
SMT+RR	0.2678	0.3858	0.2332	0.6610	0.4415
SMT+Score	0.2269	0.3407	0.2126	0.6527	0.3506
GMT+RR	0.5717	0.6178	0.556	0.7139	0.8345
GMT+Score	0.5063	0.5671	0.5178	0.7091	0.8002
mDPR+RR	0.4935	0.5223	0.5163	0.6493	0.8394
mDPR+Score	0.4852	0.5142	0.4462	0.6452	0.8418
mDPR	0.4452	0.5031	0.4462	0.7653	0.7970
KD-Encoder	0.5931	0.6058	0.5730	0.7673	0.8805
KD-SPD	**0.6265**†‡	**0.6316**†‡	**0.6049**†‡	**0.7904**†‡	**0.8912**†

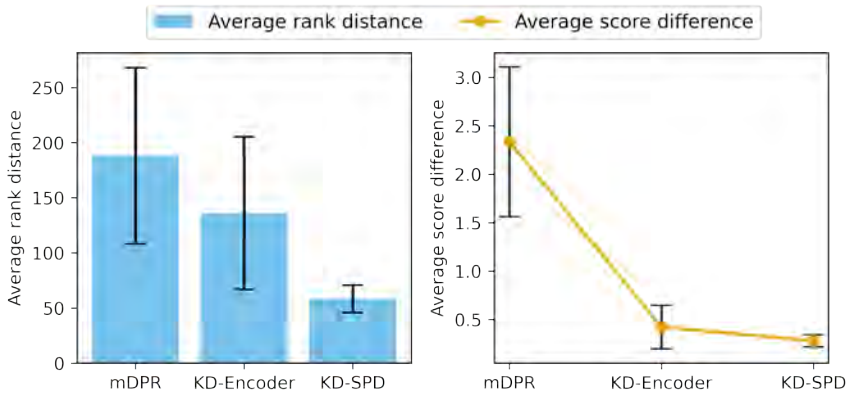


Figure 5.4: Parallel document analysis for MLIR models.

mDPR. There are two plausible reasons. First, like other multilingual models, mDPR is based on a multilingual pre-trained language model and also inherits the language bias in the pre-training step. Second, the fine-tuning steps of mDPR only focus on ranking documents within the same language space. These two reasons cause the ranking score generated by mDPR to be inconsistent across languages. Moreover, KD-Encoder performs better than mDPR on mTREC and LAReQA. Such results suggest that mapping parallel text from different languages to the same location in the vector space via knowledge distillation can efficiently transfer monolingual retrieval knowledge to multilingual settings. Finally, with the support of soft prompt decoding, KD-SPD achieves the best retrieval performance among all compared methods. In terms of precision-oriented metrics, it consistently and significantly outperforms both mDPR and KD-Encoder.

5.5.2 Analysis of Knowledge Distillation

To study how SPD behaves after knowledge distillation, we compare the rank distance and score difference of parallel relevant documents in the rank lists generated by different models. In this experiment, again, we take advantage of parallel translations in mTREC and build *duplicate* relevant documents in four languages. Thus, for each query, there are semantically similar relevant documents in different

Table 5.3: Ablation I: Decoder architecture. The numbers in the bracket show differences in percentage to KD-Encoder.

Model	**Parameter Size**	**CLEF**				
		mAP	nDCG@10	P@10	MRR	R@100
KD-Encoder	278.6M	0.1973	0.3883	0.3594	0.5641	0.4315
KD-SPD	320.0M (+14.8)	0.2200 (+11.5)	0.4160 (+7.1)	0.3714 (+3.3)	0.6356 (+12.7)	0.4689 (+8.7)
KD-UTSPD	284.5M (+2.1)	0.2075 (+5.2)	0.4023 (+3.6)	0.3722 (+3.6)	0.5964 (+5.7)	0.4576 (+6.0)

languages. Given a query, we locate all parallel relevant documents in four languages within the top 1,000 candidates from rank lists generated by mDPR, KD-Encoder, and KD-SPD, respectively. Then we compute the maximum rank distance and score difference among the four parallel documents. The equation to compute the score difference is as follows:

$$\mathcal{S} = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{1}{|\mathcal{R}_{q_i}|} \sum_{d_{kj} \in \mathcal{R}_{q_i}} \Big(\max_{k \in \boldsymbol{Y}} f'(q_i, d_{kj}) - \min_{k \in \boldsymbol{Y}} f'(q_i, d_{kj}) \Big)$$

where $\mathcal{Q}$ is the query set, $\mathcal{R}_{q_i}$ is the set of relevant documents for the query q_i, and $\boldsymbol{Y}$ is the language set. The averaging rank distance can also be obtained in a similar way. Figure 5.4 shows the results averaged over 54 queries in mTREC. We can see that KD-SPD has the smallest rank distance and score difference over parallel documents. The rank and score of parallel documents reflect the language bias in MLIR models. Thus, KD-SPD is less biased toward languages when ranking documents from a multilingual collection. Moreover, because the query embedding is fixed given the same query, the low mean and standard deviation values indicate that KD-SPD is able to generate similar embeddings for parallel documents in different languages. This matches the model design purpose.

Table 5.4: Ablation II: Effect of Teacher model. Significance tests with respect to KD-SPD (ANCE) are marked in ▲.

Teacher	CLEF					LAReQA				
	mAP	nDCG@10	P@10	MRR	R@100	mAP	nDCG@10	P@10	MRR	R@100
ANCE	0.2200	0.4160	0.3714	0.6356	0.4689	0.6265	0.6316	0.6049	0.7904	0.8912
coCondenser	0.2487▲	0.4546▲	0.4008▲	0.6826▲	0.4976▲	0.6501▲	0.6694▲	0.6436▲	0.8012	0.9172

5.5.3 Ablation Study of Decoder

In this section, we conduct experiments on two aspects that could affect the performance of KD-SPD: the number of layers in the decoder and the choice of the teacher model for distillation.

Decoder architecture. Following the idea of weights share in Transformers Dehghani et al. (2019); Jaegle et al. (2021), we replace the multi-layer (6-layer) decoder with a recurrent decoder block. Instead of N distinct layers, a decoder block has the same architecture as one decoder layer and is called recurrently for $N = 12$ steps. The weights of a decoder block are shared between steps. After each step, we add a temporal embedding $\boldsymbol{\tau} \in \mathbb{R}^{l \times d}$ to the hidden states.

$$\mathbf{H}_{d_k}^{n+1} = \boldsymbol{\tau}_n + \text{DecoderBlock}(\mathbf{H}_{d_k}^{n}, \mathbf{T}_{d_k})$$

This approach significantly reduces the size of model parameters. Named universal transformer-based SPD (UTSPD), Table 5.3 shows its performance, compared to KD-Encoder and KD-SPD. We can see that only with 2.1% more parameters, KD-UTSPD performs better than KD-Encoder. By reducing the parameter size, we show that the performance gain in SPD mainly relies on the prompt design and decoder component based on the cross-attention module. Because reducing parameters limits the model's generalization ability, there is a performance drop from distinct layers to shared weights.

Table 5.5: Zero-shot evaluation of KD-SPD. Significance tests are marked by † (over mDPR) and ‡ (over KD-Encoder).

(a) Zero-shot CLIR: English-to-Finnish.

Retrieval Method	**CLEF Finnish**				
	mAP	nDCG@10	P@10	MRR	R@100
SMT	0.0739	0.1179	0.0900	0.1390	0.1828
GMT	0.1613	0.2562	0.1560	0.4591	0.4251
mDPR	0.1682	0.2143	0.1300	0.3095	0.5010
KD-Encoder	0.1845	0.2796	0.1920	0.4537	0.5237
KD-SPD	**0.2286**†‡	**0.3321**†‡	**0.2220**†‡	**0.5092**†‡	**0.5958**†‡

(b) Zero-shot MLIR: A collection of German, Italian, and Finnish documents.

Retrieval Method	**CLEF DE-IT-FI**				
	mAP	nDCG@10	P@10	MRR	R@100
SMT+RR	0.1099	0.2245	0.208	0.4096	0.2909
SMT+Score	0.1269	0.2242	0.218	0.3726	0.2974
GMT+RR	0.1263	0.2748	0.254	0.5039	0.3384
GMT+Score	0.1447	0.2806	0.258	0.5101	0.344
mDPR+RR	0.1481	0.2734	0.268	0.391	0.3974
mDPR+Score	0.1728	0.3002	0.282	0.4816	0.4083
mDPR	0.1952	0.3377	0.306	0.5175	0.4107
KD-Encoder	0.1963	0.4262	0.382	0.6753	0.4152
KD-SPD	**0.2174**†‡	**0.4494**†‡	**0.4100**†‡	**0.7099**†‡	**0.4545**†‡

Teacher model The teacher model bounds the retrieval performance of KD-SPD. We hypothesize that a better teacher model in the English domain can lead to a better SPD model for MLIR task. Based on the leaderboard of MS MARCO passage ranking, we replace ANCE with coCondenser (Gao and Callan, 2022) for knowledge distillation. To be consistent with coCondenser, we also change the pre-trained multilingual language model used in SPD from XLM-R to mBERT. The evaluation of SPD trained with different teacher models is shown in Table 5.4. In general, KD-SPD learned from coCondenser performs better than the one learned from ANCE. This suggests that improvements with respect to the retrieval performance in the English domain can be transferred to MLIR task via KD-SPD.

5.5.4 Zero-shot Transfer

We explore the zero-shot ability of KD-SPD. For documents in languages that are not observed in the training data, we first define the language-specific vectors by averaging all trained language-specific vectors from known languages. Then KD-SPD follows the same steps as other languages to generate a prompt matrix for the new language. Thus, the knowledge learned from observed languages will be passed to the new language using the shared prompt matrix.

In this study, we focus on Finnish as the target language and use a collection of 54,694 Finnish documents from the CLEF dataset. We use this language because Finnish, a member of the Uralic language family, is distinct from the 15 languages used in training. Among the 133 English queries in the CLEF dataset, 50 have relevant annotations in the Finnish collection, forming a new set of test queries. We highlight that no training data from Finnish was used. The results in Table 5.5a show the performance of KD-SPD in CLIR task between English and Finnish, and we observe that KD-SPD significantly outperforms other methods, demonstrating the transferability of knowledge from the prompt matrices to new languages. Next, we expand the evaluation to a more challenging multilingual setting, combining Finnish with German and Italian. The resulting collection contains both observed and unobserved languages. Table 5.5b shows KD-SPD's zero-shot performance in the MLIR setting, where it still achieves the best results. This highlights KD-SPD's strong ability to transfer knowledge in a zero-shot scenario.

5.6 Summary

In this chapter, we presented a knowledge distillation (KD) framework based on soft prompt decoding (SPD) to address the language bias issue in the MLIR task. Using the soft prompt matrix as a task indicator, KD-SPD can implicitly translate documents from multiple languages into the same embedding space as the query lan-

guage. We proposed prompt decomposition to enable efficient knowledge sharing across all target languages. Our knowledge distillation framework transfers knowledge from a well-trained English retrieval model to KD-SPD, greatly reducing the data requirements for building MLIR models. Experimental results on three qualitatively different MLIR evaluation datasets show that KD-SPD significantly outperforms other baselines. Further analysis demonstrates that KD-SPD has less language bias and better zero-shot transfer ability toward new languages.

Language, while serving as a carrier for retrieval knowledge during model training, can also become a barrier for models to function across different linguistic settings. In the previous two chapters, we introduced methods for transferring retrieval knowledge across multiple languages by utilizing parallel corpora. In the next chapter, we explore the idea of removing linguistic knowledge from the embedding space. We will focus on constructing language-agnostic retrieval models through a language concept erasure task during the training to minimize the influence of language-specific information.

CHAPTER 6

LANGUAGE CONCEPT ERASURE IN DENSE RETRIEVAL MODELS

When fine-tuning PLMs for specific downstream tasks, language inherently plays a critical role as the carrier of information. Typically, the task knowledge that the model learns from the training data is heavily affected by the language of that data (Joshi et al., 2020). Consequently, the model tends to exhibit a performance bias toward the training language, often resulting in suboptimal performance on the same task in unseen languages (Nooralahzadeh et al., 2020). This limitation also applies to building retrieval models. In theory, a document's relevance to a query should transcend language barriers in most cases, excluding language-specific information needs. In practice, however, the training typically encapsulates relevance in a limited set of languages. For languages not covered in the training data, the model has to retrieve documents in the so-called zero-shot manner, leading to a performance gap between the observed and non-observed languages.

In Figure 6.1, we demonstrate this performance gap by leveraging parallel queries and documents in 11 languages from the LAReQA dataset (Roy et al., 2020). Given the same retrieval task, we gradually increase the number of possible languages used in queries and documents from 1 (English-only) to 11 languages. At each stage, we randomly assign a language from all possible languages to each query and document. We evaluate three dense retrievers with different backbone encoders that are all fine-tuned using MS MARCO passage ranking dataset (Nguyen et al., 2016). We observe that while retrieval performance in English-only settings is competitive, the

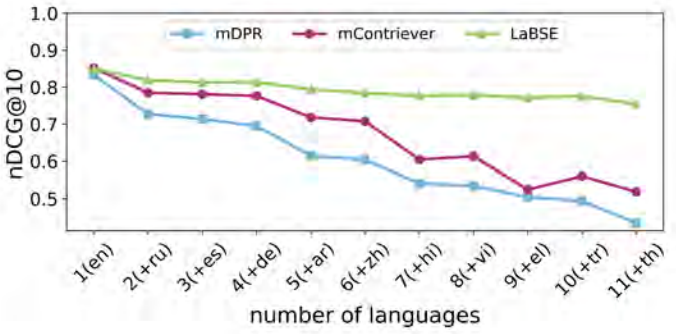


Figure 6.1: nDCG@10 decreases while the number of languages used in queries and documents increases. Results based on parallel data from LAReQA.

introduction of additional languages results in performance reductions across all models. Specifically, mDPR (Zhang et al., 2021), which is based on mBERT, shows the most sensitivity to multilingualism. In contrast, mContriever (Izacard et al., 2022) performs better due to its contrastive pre-training across multiple languages. The most robust encoder is LaBSE (language-agnostic BERT sentence embedding) (Feng et al., 2022), benefiting from its pre-training on translation ranking tasks that utilize millions of parallel sentences.

In Chapters 4 and 5, we learned that retrieval knowledge is distinct from linguistic knowledge and can be transferred across different languages. Utilizing parallel corpora, we facilitate such transfer through knowledge distillation frameworks between multiple languages. Inspired by the separation of different types of knowledge during multilingual retrieval modeling, we now explore a universal search engine that can effectively retrieve relevant information across all linguistic contexts, including monolingual (in many languages), cross-lingual, and multilingual. Different from transferring the retrieval knowledge from a well-trained English model into models for target languages, this chapter shifts focus to the training phase of dense retrieval models. We minimize the influence of language-specific information from the training data to enhance the model's ability to learn language-agnostic retrieval knowledge. In the paradigm of natural language understanding (NLU), this idea is similar to em-

bedding disentanglement (Tiyajamorn et al., 2021; Wu et al., 2022), where sentence embeddings are viewed as the combination of semantic (meaning) embedding and language-specific (syntax or idioms) embedding. For a particular task, it is preferable to concentrate model training on the semantic embedding while deliberately excluding the language-specific part. This strategy aims to enhance the model's ability to function effectively across diverse linguistic environments, improving its universal applicability and performance.

In this Chapter, we introduce Language Concept Erasure for Language-Agnostic Dense Retrieval (LANCER), a multi-task learning framework designed to reduce linguistic influence within the representation space of dense retrieval models. Given multilingual inputs, we consider language as a predictable concept tied to each input instance and design a concept erasure task to obscure the language labels within the output representations. More specifically, borrowing the concept of *guardedness* from Ravfogel et al. (2023), we first demonstrate that a zero cross-covariance matrix between features and certain labels prevents any linear classifier from detecting those labels. Based on this foundational understanding, we calculate a cross-correlation matrix between the vectors produced by a dense retriever and the language labels for each training batch. By minimizing the mean correlation values across batches, we enhance the model's language agnosticism.

While the primary task is learning retrieval knowledge, language concept erasure serves as an auxiliary task to drive the model toward generating language-agnostic representations. Concurrently, the retrieval task helps to prevent trivial solutions in the concept erasure task by ensuring that the model maintains a meaningful representation space throughout the training process. Since learning is less affected by the language of data, our method only employs retrieval data in English, such as MS MARCO, for retrieval learning. For the concept erasure task, since the language label is an inherent attribute tied to any context of a certain language, LANCER is

Table 6.1: Language identification accuracy of logistic regression on mPLMs and retrieval models. Train test splits are sampled from mC4 dataset.

1. Input features are taken from mPLMs.				
Models	mBERT	XLM-R	mT5	mE5
Accuracy	98.1	96.1	97.2	98.0
2.Input features are taken from dense retrievers.				
Models	mDPR (mBERT)	mDPR (mT5)	mContriever	LaBSE
Accuracy	97.9	96.7	98.0	81.6

capable of operating on multilingual non-parallel corpora, such as mC4 (Xue et al., 2021) and Wiki-40B (Guo et al., 2020), effectively diminishing the necessity for parallel data toward language-invariant modeling. The dense retrieval models developed using our framework exhibit reduced language bias and show consistent improvements on monolingual, cross-lingual, and multilingual retrieval tasks.

The work described in this chapter, namely Language Concept Erasure for Language-invariant Dense Retrieval, is under submission. I was the lead author who designed the model architecture and conducted the experiments.

6.1 Language Agnostic via Concept Erasure

Our objective is to diminish the linguistic features within the embedding space of a dense retrieval model, enabling it to produce language-agnostic representations across various linguistic contexts. To achieve this, we structure our goal into a training task of hiding language labels from any linear classifier. In this section, we first measure the language-specific signals by a language identification task using logistic regression classifiers. Then, we borrow the definition of linear guardedness from Ravfogel et al. (2023) and convert its equivalent assumption proved by Belrose et al. (2024) to a loss function for language concept erasure. Finally, we present the multi-task learning framework, LANCER, and the details of training a dense retrieval model.

6.1.1 Language Identification

In most search scenarios, the relevance between a query and a document should transcend the language in which they are expressed (Rahimi et al., 2020; Zhang and Zhao, 2020). For instance, if a document is labeled relevant to a query in English, translating both the query and the document into Arabic should not alter the judgment of their relevance. This principle highlights the importance of language-independent factors in determining the relevance of search results.

However, we find that mPLMs and the retrieval models built upon them still retain strong language-specific signals in their output representations. As shown in Table 6.1, we use the dense vectors from 16 languages as input features to train a logistic regression classifier for predicting language labels. The high accuracy achieved on a held-out test set indicates that the language factor has a strong influence on the dense vector used for relevance scoring. This suggests that despite the intent to transcend linguistic boundaries, current models still embed significant language-specific information, which could impact their effectiveness in information retrieval tasks across diverse linguistic settings.

6.1.2 Language Concept Erasure

Based on this observation, we propose language concept erasure to reduce the influence of language in relevance scoring. Specifically, our goal is to prevent any linear classifier from detecting the language label given the dense vectors, thereby largely erasing the identity of the source language. We adopt the idea of **guardedness** from Ravfogel et al. (2023) to formally define the task of language concept erasure.

Given vector $\mathrm{X} \in \mathbb{R}^k$, and a concept Z (the one-hot labels) taking values in $\mathcal{Z} = \{\mathbf{z} \in \{0,1\}^n \mid \|\mathbf{z}\|_1 = 1\}$. Let $\mathcal{V} = \{\eta(\cdot;\boldsymbol{\theta}) : \mathbb{R}^k \to \mathbb{R}^n \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ be the class of all linear predictors, taking the form $\eta(\mathrm{X}) = \mathbf{W}\mathrm{X} + \mathbf{b}$ for some weight matrix $\mathbf{W} \in \mathbb{R}^{n \times k}$

and bias $\mathbf{b} \in \mathbb{R}^n$, We say X *linearly guards* Z if no classifier in $\mathcal{V}$ can outperform a constant function at predicting Z.

Next, we prove that linear guardedness is achieved when the class-conditional mean equals the unconditional mean.

Theorem 6.1 Suppose $\mathcal{L}$ is convex loss functions defined on $(\eta(\mathrm{X}), \mathrm{Z})$. If each class-conditional mean $\mathbb{E}[\mathrm{X}|\mathrm{Z} = j]$ is equal to $\mathbb{E}[\mathrm{X}]$, then the constant predictor cannot be improved upon.

Proof. By Jensen's inequality, the loss with η evaluated on X is lower bounded by the loss with η evaluated on the unconditional mean of the data.

$$\begin{aligned}
\mathbb{E}\big[\mathcal{L}(\eta, \mathrm{Z})\big] &= \mathbb{E}_{\mathrm{Z}}\big[\mathbb{E}\big[\mathcal{L}(\eta, \mathrm{Z})\big]\big] \\
&\geq \mathbb{E}_{\mathrm{Z}}\big[\mathcal{L}(\mathbb{E}[\eta|\mathrm{Z}], \mathrm{Z})\big] \quad \text{(By Jensen's inequality)} \\
&= \mathbb{E}_{\mathrm{Z}}\big[\mathcal{L}(\mathbf{W}\mathbb{E}[\mathrm{X}|\mathrm{Z}] + \mathbf{b}, \mathrm{Z})\big] \quad \text{(By linearity of } \eta) \\
&= \mathbb{E}_{\mathrm{Z}}\big[\mathcal{L}(\mathbf{W}\mathbb{E}[\mathrm{X}] + \mathbf{b}, \mathrm{Z})\big] \quad \text{(By assumption)}
\end{aligned}$$

This represents the loss from a constant predictor $\eta' = \mathbf{W}\mathbb{E}[\mathrm{X}] + \mathbf{b}$. Since every predictor's loss is lower bounded by a constant predictor, X linearly guards Z. □

Intuitively, this shows that the expected loss of the classifier is lower-bounded by loss calculated from the class centroids. When all centroids are identical, the minimal achievable loss corresponds to replacing every data point with the global mean $\mathbb{E}[\mathrm{X}]$. Therefore, to achieve linear guardedness, we need to drive the class-conditional mean to the unconditional mean. Next, we prove that Theorem 6.1 is equivalent to zero covariance between every component of X and every component of Z.

Theorem 6.2 Let X and Z following the settings above, and each class probability $P(\mathrm{Z} = j)$ nonzero. Then the class-conditional means $\mathbb{E}[\mathrm{X}|\mathrm{Z} = j]$ are all equal to the unconditional mean $\mathbb{E}[\mathrm{X}]$ if and only if the cross-covariance matrix Σ_{XZ}, whose $(i, j)^{th}$ entry is $\mathrm{Cov}(\mathrm{X}_i, \mathrm{Z}_j)$, is a zero matrix.

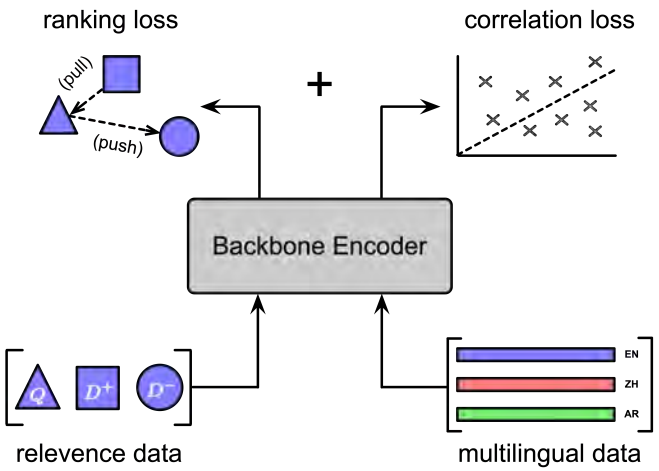


Figure 6.2: LANCER training objectives.

Proof. We can rewrite the $(i, j)^{th}$ entry of the cross-covariance matrix Σ_{XZ} as:

$$\mathbb{E}[\mathrm{X}_i\mathrm{Z}_j] - \mathbb{E}[\mathrm{X}_i]\mathbb{E}[\mathrm{Z}_i] = P(\mathrm{Z} = j)\big(\mathbb{E}[\mathrm{X}_i|\mathrm{Z} = j] - \mathbb{E}[\mathrm{X}_i]\big)$$

Since $P(\mathrm{Z} = j) > 0$, then $\mathbb{E}[\mathrm{X}_i|\mathrm{Z} = j] = \mathbb{E}[\mathrm{X}_i]$ if and only if $\mathbb{E}[\mathrm{X}_i\mathrm{Z}_j] = \mathbb{E}[\mathrm{X}_i]\mathbb{E}[\mathrm{Z}_i]$, which is $\mathrm{Cov}(\mathrm{X}_i, \mathrm{Z}_j) = 0$ □

We finally establish a concrete condition for linear guardedness. Suppose X is the multilingual dense vector generated by the backbone encoder and Z is the language labels tied to input instances. We define language concept erasure as being equivalent to X linearly guards Z. Thus, if (X, Z) satisfies Theorem 6.2, then the dense retrieval model prevents any linear classifier from detecting languages from its outputs.

6.1.3 Multi-task Learning

In practice, Theorem 6.2 is a very weak condition. It does not identify a unique solution of X. In fact, any trivial vector satisfies the condition of zero cross-covariance. However, we can convert Theorem 6.2 into a loss function and pair it with the ranking loss to form a multi-task learning framework. During training, the model takes two types of data inputs in each batch:

(i) Retrieval data, which includes triplets consisting of queries (Q), positive documents (D^+), and negative documents (D^-). We calculate the ranking loss, L_R, through contrastive learning Chen et al. (2020):

$$\mathcal{L}_R = \sum_{q \in Q} \sum_{d^+ \in D^+} -\log \frac{e^{s(q,d^+)}}{e^{s(q,d^+)} + \sum\limits_{d^- \in D^-} e^{s(q,d^-)}}$$

(ii) Multilingual data, which is a group of passages (p) with language label (z), $\{(p_j, z_i)_{j=1}^m\}_{i=1}^n$, where n is the number of languages and m is the number of passages per language. We sample the same number of passages from each language per batch to ensure languages are equally represented. We compute the cross-covariance matrix between dense vectors of input passages $\mathbf{X}$, and language labels $\mathbf{Z}$.

$$\Sigma_{\mathbf{XZ}} = \mathbb{E}[(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{Z} - \bar{\mathbf{Z}})^\top]$$

The scale of the embedding values significantly influences the magnitude of covariance. Unnormalized outputs from some encoders result in covariance values that vary widely across different input instances. Therefore, we standardize the covariance matrix into the correlation matrix by dividing by the standard deviations: $\rho_{\mathbf{XZ}} = \Sigma_{\mathbf{XZ}} / \sigma_{\mathbf{X}} \sigma_{\mathbf{Z}}$. The concept erasure loss is defined as the mean absolute value of the correlation matrix:

$$\mathcal{L}_C = \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} |corr(\mathrm{X}_i, \mathrm{Z}_j)|$$

By Theorem 6.2, we aim for this value to be as close to zero as possible. These two types of data and their corresponding losses complement each other effectively. The concept erasure task effectively removes language-specific information from the dense vectors, enabling the retrieval task to concentrate on language-agnostic knowledge. Simultaneously, the retrieval task, which focuses on semantic matching, ensures that the model maintains meaningful representations throughout the training. This

balance prevents the concept erasure task from degenerating. Finally, as shown in Figure 6.2, we add the primary ranking loss $\mathcal{L}_R$ for retrieval and the correlation loss for concept erasure $\mathcal{L}_C$ to conduct the training of dense retrievers.

$$\mathcal{L} = \mathcal{L}_R + \mathcal{L}_C$$

Training data requirement. Because the language label is an intrinsic attribute separate from semantic meaning, the language concept erasure task of LANCER has minimal data requirements for multilingual input. A clean corpus from each language is sufficient to support the running of this task. On the other hand, as the retrieval task is less influenced by language-specific information, it can utilize query-document pairs in any language. In experiments, we focus on retrieval data in English for training.

6.2 Experimental Setup

6.2.1 Modeling Details

We apply LANCER to multilingual dense retrieval models with different degrees of multilingual pre-training: mBERT (mDPR), mContriever (Izacard et al., 2022), and LaBSE (Feng et al., 2022). The pre-training of mBERT is only an extension of Masked Language Modeling (MLM) and next sentence prediction (NSP) to 104 languages. Based on mBERT, mContriever is further pre-trained on unsupervised contrastive learning over 29 languages. LaBSE, also built on mBERT, is further pre-trained on the translation ranking task, leveraging millions of parallel text. To compare with existing baselines, we use the MS MARCO passage ranking dataset (Nguyen et al., 2016) as the retrieval training data. Note that there is no existing LaBSE-based dense retriever built on MS MARCO, so we created one by fine-tuning LaBSE on MS MARCO. For language concept erasure, we use a multilingual corpus containing 16

languages[1]. We sample 3M passages per language from the mC4 (Xue et al., 2021) dataset. We set the batch size to 64 for the retrieval task and 256 (16 examples per language) for the concept erasure task. Based on the size of the MS MARCO train split, we train each model for 4 epochs with a learning rate of $2e^{-5}$.

To evaluate language-agnostic dense retrievers refined by LANCER, we conduct experiments on various benchmark retrieval datasets covering multilingual, cross-lingual, and monolingual (in many languages) tasks. Some of the evaluation datasets include training splits. To assess language agnosticism, we do not perform any additional fine-tuning using those splits to keep zero-shot evaluations of our approach and all compared methods.

6.2.2 Datasets and Metrics

6.2.2.1 Multilingual

CLEF. We evaluate searching a multilingual collection using English queries. This dataset is reused from Section 5.4.1 to evaluate one-to-many setting of MLIR.

LAReQA. We evaluate the retrieval performance when the query and collection are both multilingual. LAReQA (Roy et al., 2020) is a benchmark for language-agnostic answer retrieval from a multilingual candidate pool. Different from Section 5.4.1, we evaluate many-to-many setting of MLIR by including queries from 11 languages.

6.2.2.2 Cross-lingual and Monolingual

XOR-Retrieve. We evaluate searching English collection using queries in other languages. XOR-Retrieve (Asai et al., 2021) is a benchmark for evaluating cross-lingual retrieval systems. It includes 7 cross-lingual tasks between target language queries and English documents. The corpus contains 18.2M passages with a maximum of 100 word tokens from the English Wikipedia.

---

[1]List of training languages (ISO code): ar, bn, de, en, es, fa, fi, fr, hi, id, ja, ko, ru, te, th, zh

XTREME-UP. XTREME-UP (Ruder et al., 2023) focuses on extremely low-resource languages. Similar to XOR-Retrieve, it includes 20 cross-lingual tasks of queries in low-resource language and documents in English.

MIRACL. We evaluate monolingual retrieval across multiple languages. MIRACL (Zhang et al., 2023) has a broad language coverage for evaluating monolingual retrieval. Developed on top of Mr.TyDi (Zhang et al., 2021), MIRACL comprises data in 18 languages, with both queries and documents presented in the same language.

6.2.2.3 Metrics

We report mAP and nDCG@10 for both multilingual evaluation datasets (CLEF and LAReQA). To be consistent with the measures previously reported on the benchmark datasets (Li et al., 2022; Ruder et al., 2023; Zhang et al., 2023), we report nDCG@10 on MIRACL and MRR@10 on XTREME-UP. For XOR-Retrieve, we evaluate recall on the first 5,000 tokens retrieved, denoted as Recall@5kt. We determine statistical significance using the two-tailed paired t-test with p-value less than 0.05 (i.e., 95% confidence level).

6.2.3 Compared Methods

Across all evaluations, we compared the performance of models incorporating LANCER to those without, e.g., mDPR+LANCER vs. mDPR. For multilingual evaluation, we included the following additional baselines for comparison:

LSAR: As an unsupervised method, LSAR (Xie et al., 2022) is based on matrix decomposition to identify a language-agnostic subspace and then directly projects the original multilingual embeddings onto that subspace to reduce the effects of language on downstream tasks.

LEACE: Also worked as an unsupervised method, LEACE (Belrose et al., 2024) derives a projection in closed-form to prevent linear classifiers from detecting a concept. We apply it upon baseline retrievers to reduce the effects of language concepts.

Table 6.2: Results for multilingual retrieval on CLEF and LAReQA. LAReQA (Full) includes parallel queries and documents in 11 languages. LAReQA (Sampled) refers to randomly selecting a language for each query and document. Results are averaged over five folds. Our approaches are *highlighted* in light blue with significant improvements marked by † (over LEACE), ‡ (over KD-SPD), and ◇ (over baseline model).

Method	CLEF		LAReQA (Full)		LAReQA (Sampled)	
	mAP	nDCG@10	mAP	nDCG@10	mAP	nDCG@10
KD-SPD	22.0	41.6	48.4	50.4	55.5	60.0
mDPR	20.2	34.6	25.5	31.7	41.0	41.6
+ LSAR	19.8	35.8	34.0	39.2	48.9	53.3
+ LEACE	18.9	34.6	33.4	38.7	48.9	53.2
+ LANCER	21.6†	39.1$^{\dagger\diamond}$	39.3$^{\dagger\diamond}$	43.3$^{\dagger\diamond}$	53.1$^{\dagger\diamond}$	57.7$^{\dagger\diamond}$
mContriever	27.2	46.1	31.1	37.3	48.8	52.5
+ LSAR	26.9	47.4	38.8	43.8	55.8	60.2
+ LEACE	28.3	48.8	39.1	44.2	56.3	60.7
+ LANCER	30.0$^{\dagger\ddagger\diamond}$	50.7$^{\dagger\ddagger\diamond}$	42.6$^{\dagger\diamond}$	47.6$^{\dagger\diamond}$	58.4$^{\dagger\ddagger\diamond}$	62.8$^{\dagger\ddagger\diamond}$
LaBSE	24.0	44.2	62.9	64.0	72.4	76.2
+ LSAR	22.8	42.5	61.4	62.1	70.9	74.9
+ LEACE	23.9	44.6	61.2	62.0	71.3	75.2
+ LANCER	25.8†‡	47.0$^{\dagger\ddagger\diamond}$	64.5$^{\dagger\ddagger\diamond}$	65.2†‡	74.5$^{\dagger\ddagger\diamond}$	78.1$^{\dagger\ddagger\diamond}$

KD-SPD: Based on knowledge distillation, KD-SPD (Huang et al., 2023b) designed a language-aware decomposition prompt for the encoder to transfer knowledge from an English retriever to multiple languages using parallel corpora.

For cross-lingual and monolingual tasks, we include results from **SWIM-X** (Thakur et al., 2023), a synthetic query generation method using LLMs. It utilizes in-domain documents to generate synthetic queries and then performs fine-tuning to build multilingual dense retrieval models.

6.3 Results

6.3.1 Retrieval Performance

Multilingual. Table 6.2 lists the multilingual evaluation results. We observe that when LANCER is applied, all three baseline models show substantial improvements on two datasets in terms of both mAP and nDCG@10. Note that retrieval data used for training remained consistent across these experiments. Because of the language

Table 6.3: Results showing Recall@5kt (%) for cross-lingual retrieval on XOR-Retrieve dev (labels of test split are not released). WR denotes the win ratio of LANCER over baseline.

Method	Avg.	ar	bn	fi	ja	ko	ru	te	WR
SWIM-X	59.0	54.0	67.4	59.2	52.7	55.1	54.4	70.2	-
mDPR	39.3	34.3	35.5	45.2	40.2	36.5	43.9	39.5	-
+ LANCER	41.4	36.2	37.8	47.1	37.8	45.3	42.2	43.3	5/7
mContriever	44.0	37.5	38.2	50.6	41.1	37.2	49.8	53.8	-
+ LANCER	45.7	43.0	35.9	56.4	39.4	46.0	43.5	55.5	4/7
LaBSE	56.8	56.0	63.5	57.6	50.2	50.2	48.1	71.8	-
+ LANCER	57.2	54.4	62.5	58.3	51.0	52.6	47.3	74.4	4/7

Table 6.4: Results showing MRR@10 (%) for cross-lingual retrieval on XTREME-UP test. WR denotes the win ratio of LANCER over baseline.

Method	Avg.	as	bho	brx	gbm	gom	gu	hi	hne	kn	mai	ml	mni	mr	mwr	or	pa	ps	sa	ta	ur	WR
SWIM-X	25.2	24.4	27.7	4.3	28.3	25.4	29.4	32.4	28.8	30.1	31.8	34.4	5.1	30.7	25.7	15.8	29.6	20.6	26.1	27.9	26.1	-
mDPR	5.9	2.6	6.5	0.6	7.0	2.2	5.4	13.9	5.7	6.3	6.9	8.7	0.3	8.7	6.1	0.7	9.5	2.6	4.1	7.7	13.3	-
+LANCER	9.8	5.0	9.5	0.8	11.2	6.3	11.9	19.2	10.0	10.5	11.5	14.1	0.8	15.9	9.9	0.3	15.5	3.7	9.5	13.8	16.4	19/20
mContriever	4.6	3.6	5.4	0.9	6.3	1.8	2.2	10.9	5.3	5.5	7.0	4.3	0.9	6.1	6.6	0.8	5.3	2.0	4.4	7.9	5.7	-
+LANCER	6.5	5.1	6.4	1.0	9.7	3.3	4.2	13.4	7.4	8.8	9.0	6.3	0.7	9.3	8.8	0.7	6.9	3.0	8.4	8.0	8.5	18/20
LaBSE	28.3	25.0	28.3	2.8	29.4	21.0	36.2	38.5	27.6	36.3	31.9	36.9	4.5	37.9	28.6	27.0	35.5	22.2	27.4	35.6	34.1	-
+LANCER	29.2	26.1	29.2	2.4	27.4	22.5	37.7	40.7	26.2	38.9	31.7	38.5	4.1	39.0	28.4	29.6	36.5	22.9	28.1	37.3	36.3	14/20

concept erasure, models built with LANCER have less language bias, leading to better performance on multilingual tasks.

Moreover, LANCER outperforms post-hoc methods (LSAR and LEACE). Compared with the knowledge transfer method, LaBSE+LANCER uniformly improves KD-SPD, while mContriever+LANCER also performs better except on LAReQA (Full). Lastly, from a task perspective, LAReQA presents a greater challenge than CLEF due to the inclusion of more languages in its queries and documents. Because LaBSE is pre-trained on a wide range of languages using parallel sentences, mContriever is able to surpass LaBSE on CLEF but falls behind on LAReQA.

Cross-lingual. Table 6.3 and Table 6.4 list cross-lingual results on XOR-Retrieve and XTREME-UP respectively. On XOR-Retrieve, LANCER demonstrates competitive performance compared to corresponding baseline models, improving 2.1 points on mDPR and 1.7 points on mContriever. When applied to LaBSE, LANCER aligns

Table 6.5: Results showing nDCG@10 (%) for monolingual retrieval on MIRACL dev (labels of test split are not released). WR denotes the win ratio of LANCER over baseline.

Method	Avg.	ar	bn	en	es	fa	fi	fr	hi	id	ja	ko	ru	sw	te	th	zh	de	yo	WR
SWIM-X	46.4	60.2	57.1	34.7	33.4	36.3	40.6	64.3	33.0	39.5	40.8	43.3	49.7	40.0	55.9	56.3	63.3	60.2	57.1	-
mDPR	41.8	49.9	44.3	39.4	47.8	48.0	47.2	43.5	38.3	27.2	43.9	41.9	40.7	29.9	35.6	35.8	51.2	49.0	39.6	-
+ LANCER	47.5	55.6	50.5	43.4	46.3	49.1	56.6	46.1	36.7	34.3	49.0	47.4	46.7	39.4	52.7	46.1	50.0	46.8	58.5	14/18
mContriever	37.8	49.1	48.4	32.7	33.3	37.1	48.4	27.0	35.9	32.7	34.1	40.2	35.1	44.5	46.2	45.0	27.5	29.7	33.7	-
+ LANCER	50.1	61.4	56.9	40.7	46.1	38.0	65.4	41.2	35.7	43.6	48.1	54.5	46.2	58.0	67.9	58.2	45.1	43.2	51.7	17/18
LaBSE	45.6	50.2	53.7	35.6	37.7	42.4	57.2	40.6	41.4	37.8	34.6	46.2	40.5	57.4	53.9	50.1	34.9	39.7	67.7	-
+ LANCER	48.1	52.9	57.2	37.5	38.0	45.9	60.6	41.7	43.8	39.5	39.4	48.8	42.2	57.9	58.9	55.2	37.3	38.1	70.2	17/18

closely with the baseline. SWIM-X performs the best on XOR-Retrieve. However, SWIM-X utilizes in-domain data to generate cross-lingual training pairs, while our experiments are completely zero-shot evaluations. For collections with strong domain features like Wikipedia, synthetic data not only supports language-specific training but also acts as a form of domain adaptation, contributing to this strong performance.

On XTREME-UP, LANCER consistently enhances performance over the baseline models. Both LaBSE and LaBSE+LANCER surpass SWIM-X. The performance of SWIM-X suggests that using LLMs for data augmentation does not always yield high-quality data, particularly in low-resource languages.

Monolingual. Table 6.5 lists monolingual results on MIRACL, covering 18 languages. Compared to cross-lingual, LANCER improves the corresponding baseline models on monolingual tasks by a large margin. Specifically, in terms of nDCG@10, LANCER achieves an improvement of 5.7 points (13.6%) over mDPR, 12.3 points (32.5%) over mContriever, and 2.5 points (5.5%) over LaBSE. Surprisingly, when LANCER is applied, all three models outperform SWIM-X. This suggests that LANCER has robust zero-shot capability in monolingual tasks, highlighting its effectiveness without retrieval training for specific languages. From the data perspective, this also suggests that when language bias is reduced in embedding space, retrieval knowledge

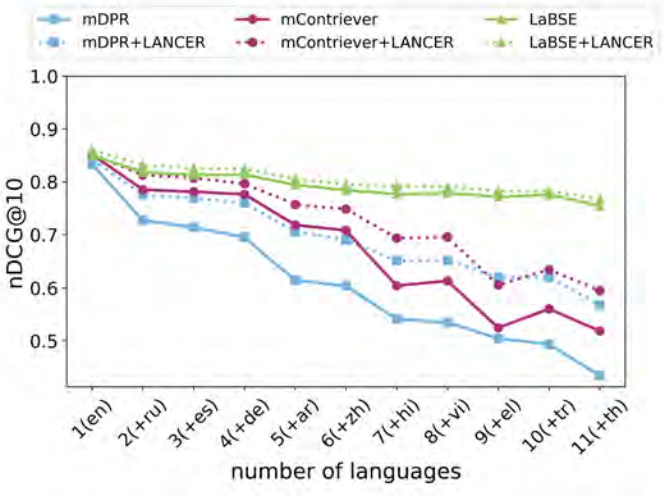


Figure 6.3: Compared to corresponding baselines, LANCER shows more robust nDCG@10 against the increase of languages. Results based on LAReQA.

provided by MS MARCO data (English) is more comprehensive than language-specific synthetic data generated by current LLMs.

6.3.2 Effect of Multilingualism

At the beginning of this chapter, as shown in Figure 6.1, we demonstrate the language bias in multilingual retrieval by increasing the number of languages used in queries and documents. Here, we replicate the experiment with models trained using LANCER, observing how their performance shifts as more languages are incorporated into the queries and documents. In Figure 6.3, the models demonstrate improved resilience to language bias as the number of languages increases, maintaining higher levels of nDCG@10 compared to those without LANCER. This suggests that LANCER effectively mitigates the challenges posed by linguistic diversity, enhancing the model's ability to handle multilingual information retrieval more robustly.

6.3.3 Analysis of Training

To study the impact of language concept erasure on the dense vectors, we leverage held-out multilingual train and test splits to monitor language label recovery. At

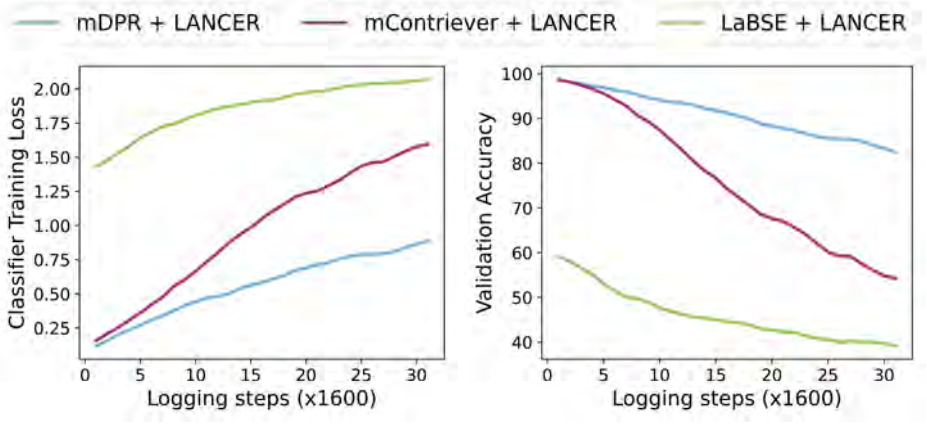


Figure 6.4: Training loss of logistic regression (Left) and prediction accuracy (Right) for language label recovery.

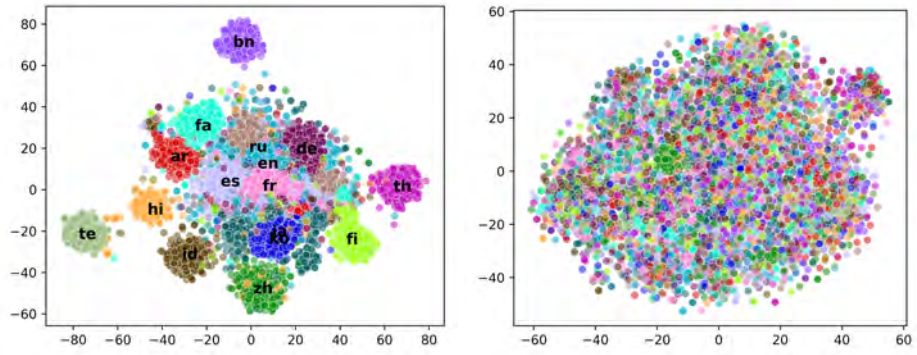


Figure 6.5: t-SNE visualization of multilingual representations from mDPR (Left) versus mDPR+LANCER (Right). Best viewed in color.

each logging step, we map the held-out splits into dense vectors and build a logistic regression classifier to predict language labels. Figure 6.4 records the loss on train split (Left) and prediction accuracy on test split (Right) on three models. As training continues, it is harder for the classifier to identify the language label according to rising loss and declining accuracy. This trend indicates that the language concept erasure task effectively reduces the language information in the dense vectors, making the model more language-agnostic.

6.3.4 Analysis of Representation

To further demonstrate the impact of LANCER, we analyze the representations produced by dense retrieval models, both with and without the language concept erasure task. We sample 300 passages per language from 16 training languages and use t-SNE to visualize their representations. In Figure 6.5, the visualizations from mDPR show that the representations are predominantly clustered by language. However, after integrating LANCER, the representations from different languages are intermingled. This further supports that LANCER effectively diminishes language-specific clustering, resulting in a more language-agnostic embedding space.

6.4 Summary

In this chapter, we introduce LANCER, a multi-task training framework designed to improve language-agnostic dense retrieval. The core of LANCER is the language concept erasure task, which reduces the language-specific signals present in the multilingual dense vectors by preventing linear classifiers from detecting the language labels. Paired with the retrieval task, LANCER enables the model to prioritize learning language-agnostic knowledge for query-document matching.

We conduct experiments across all possible linguistic settings of an IR task (e.g., monolingual, cross-lingual, and multilingual). The extensive results from these experiments demonstrate the effectiveness of LANCER in building language-agnostic dense retrieval models. In multilingual contexts, LANCER outperforms knowledge transfer using parallel data. Furthermore, in monolingual tasks across 18 languages, LANCER, as a zero-shot approach, surpasses an in-domain data augmentation method based on LLMs.

Despite the effectiveness of language concept erasure, our experimental results (e.g., Figure 6.3 and 6.4) indicate the potential for further reducing language bias and enhancing multilingual retrieval performance. Moreover, our method is limited

to linear classifiers. The language labels can still be recovered accurately by non-linear classifiers like multi-layer perceptron (MLP). Our community still has a considerable path to tread in order to overcome language bias in retrieval systems.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

Overcoming language barriers, Cross-lingual Information Retrieval (CLIR) and Multilingual Information Retrieval (MLIR) enhance the comprehensive satisfaction of users' informational needs. Generally, these tasks require matching queries and documents in different languages. In addition to the ranking component, the retrieval models for CLIR or MLIR need to possess some form of translation knowledge to map the vocabulary of the query language to that of the documents' language. The effectiveness of retrieval depends on the model's ability to match queries with documents and bridge the linguistic gap between them.

Based on the Transformer architecture, multilingual pre-trained language models (mPLMs) promote the joint learning of contextualized representations for multiple languages with the same model. Because tokens in different languages are projected into the same representation space, these models can also be adopted as the source of translation knowledge to bridge the linguistic gap. Fine-tuning these models with retrieval-specific data enables them to learn the knowledge of query-document matching and perform retrieval tasks across diverse linguistic settings (e.g., monolingual, cross-lingual, and multilingual).

In this dissertation, we focus on the challenges of building neural ranking models using mPLMs, covering cross-encoder reranker and bi-encoder dense retriever. Specifically, we study the translation gap (Chapter 3), data scarcity (Chapter 4), and language bias (Chapter 5 & Chapter 6) issues in CLIR and MLIR tasks.

First, from monolingual to cross-lingual, the words co-occurring in query and document become translations, creating difficulties for the model to catch the exact match signals. Re-introducing external translation knowledge into the CLIR models effectively reduces such translation gap. For the cross-encoder reranker, we inject word-level dictionary knowledge as a translation attention matrix into the Transformer layer, parallelizing with the multi-head attention mechanism. By improving the token similarity of mutually translated words in query and document, our design improves cross-lingual document reranking on both high- and low-resource languages.

Then, unlike the English retrieval task, which benefits from abundant resources for model training, the scarcity of retrieval data in other languages, especially in low-resource languages, makes it challenging to build CLIR models. While previous approaches mainly focused on synthesizing multilingual datasets, we explore transferring retrieval knowledge learned from English retrieval data to other languages to address the data scarcity problem. We propose a knowledge distillation framework using parallel data and cast cross-lingual token alignment as the optimal transport problem loss computation. By reducing the data requirement from cross-lingual relevance labels to parallel sentences, our method significantly improves CLIR performance involving low-resource languages.

Moreover, expanding from CLIR to MLIR, where queries and documents involve more languages, presents challenges beyond the increased need for multilingual training data. When a search collection encompasses documents in multiple languages, it requires that the model maintain consistent and fair performance across different languages. Following the idea of knowledge transfer, we develop a language-aware decomposition prompt for the encoder to transfer knowledge from an English retriever to multiple languages using parallel corpora. Our proposed method uses English as a pivot language and maps document representations from other languages into

the same English embedding space. This strategy effectively reduces language bias, thereby enhancing the performance of MLIR tasks.

Finally, based on the findings of knowledge transfer through distillation, we argue that retrieval knowledge can be separated from linguistic knowledge. Therefore, we introduce a multi-task learning framework to build language-agnostic dense retrieval models. The core design principle of this framework is to minimize the linguistic signals within the representation space. We consider language as a predictable label from model outputs and employ the condition of linear guardedness to design a loss function for language concept erasure. Dense retrieval models that incorporate language concept erasure are less sensitive to the input language and exhibit substantial improvements across a variety of retrieval tasks, including monolingual (in many languages), cross-lingual, and multilingual settings. Our approach enhances the model's ability to function effectively across diverse linguistic environments, improving its universal applicability and performance.

7.2 Future Work

In our increasingly interconnected world, as the content on the internet and digital platforms becomes more global, facilitating information access beyond the language barrier is a meaningful and important research topic. While we delved into certain aspects of CLIR and MLIR, there are still areas of interest and challenges that we have not yet addressed. Next, we will briefly outline potential directions for future research.

7.2.1 Language Coverage Expansion

We have developed tasks and methodologies to overcome data scarcity issues and expand the neural ranking models to more languages, especially underrepresented languages. We reduced the data requirement for building retrieval models from rel-

evance labels to bitext data and then to monolingual corpora in target languages. However, existing datasets are far from sufficient to fully develop information access capabilities for the 7000+ languages spoken on our planet. Existing mPLMs cover a maximum of a couple of hundred languages. The rest are extremely low-resource languages that do not even have a monolingual corpus to support the training of language models. The performance of retrieval models is greatly limited in those languages. Achieving a broad coverage of languages is still a significant challenge for both the NLP and IR communities.

As shown in Chapter 6, one promising way to overcome the limitation of languages is to build retrieval models by focusing on language-independent knowledge for query document matching. We hope our findings and methodologies can shed light on following up research for language-agnostic retrieval.

7.2.2 Query-based Language Preference

In this dissertation, our efforts to improve retrieval performance are mainly under a particular task assumption, such as cross-lingual or multilingual. The ability to assess language preferences based on individual queries has been overlooked. In many real-world applications, particularly in web search scenarios, the linguistic context often encompasses a mix of monolingual, cross-lingual, and multilingual elements.

The linguistic setting of retrieval should be evaluated on a query-to-query basis. For some queries, monolingual retrieval is enough to fulfill the user's information needs, while other queries require CLIR or even MLIR. Future research could focus on developing adaptive retrieval systems that intelligently predict the language preferences for retrieval based on the query intent and retrieve documents accordingly with appropriate models.

7.2.3 Multilingual Retrieval-augmented Generation

An extension of the MLIR task is multilingual content aggregation. We have developed methods to reduce language bias and retrieve documents fairly across languages based on their relevance to a given query. The final step to fulfill the information needs is to summarize the multilingual ranked list into the language of the query for users to understand. A straightforward approach is to apply techniques related to query-based multi-document summarization. However, the information contained in documents from different languages could be duplicated, complementary, or even contradictory. Relevance in IR does not imply correctness, completeness, or accountability. A document can be very relevant and misleading to a query simultaneously. It is important for the retrieval model to evaluate the rank of a document from more comprehensive aspects rather than only relevance.

Moreover, with the general-purpose text generation ability from Large Language Models (LLMs), we envision in the coming future, in most cases, the search results will no longer be rendered to the user directly but be used as external knowledge input along with the query for LLMs to generate a response. To support this new search paradigm, retrieval systems need to not only identify relevant documents across different languages but also generate comprehensive ranked lists, considering the source's reliability, the content's factual accuracy, and the potential biases present in the information.

REFERENCES

Adriani, M. (2000). Using statistical term similarity for sense disambiguation in cross-language information retrieval. *Information retrieval*, 2:71–82.

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Asai, A., Kasai, J., Clark, J., Lee, K., Choi, E., and Hajishirzi, H. (2021). XOR QA: Cross-lingual open-retrieval question answering. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 547–564, Online. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Ballesteros, L. and Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Forum*, volume 31, pages 84–91. ACM New York, NY, USA.

Ballesteros, L. and Croft, W. B. (1998). Resolving ambiguity for cross-language retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71.

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. (2024). Leace: perfect linear concept erasure in closed form. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Bonab, H., Allan, J., and Sitaraman, R. (2019). Simulating clir translation resource scarcity using high-resource languages. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 129–136, New York, NY, USA. Association for Computing Machinery.

Bonab, H., Sarwar, S. M., and Allan, J. (2020). Training effective neural clir by bridging the translation gap. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 9–18, New York, NY, USA. Association for Computing Machinery.

Bonifacio, L. H., Campiotti, I., Jeronymo, V., Lotufo, R., and Nogueira, R. (2021). mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.

Braschler, M. (2001). Clef 2000 — overview of results. In Peters, C., editor, *Cross-Language Information Retrieval and Evaluation*, pages 89–101, Berlin, Heidelberg. Springer Berlin Heidelberg.

Braschler, M. (2002a). Clef 2001 — overview of results. In Peters, C., Braschler, M., Gonzalo, J., and Kluck, M., editors, *Evaluation of Cross-Language Information Retrieval Systems*, pages 9–26, Berlin, Heidelberg. Springer Berlin Heidelberg.

Braschler, M. (2002b). Clef 2002—overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 9–27. Springer.

Braschler, M. (2003). Clef 2003–overview of results. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 44–63. Springer.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chaware, S. and Rao, S. (2009). Information retrieval in multilingual environment. In *2009 Second International Conference on Emerging Trends in Engineering & Technology*, pages 648–652. IEEE.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Cieri, C., Maxwell, M., Strassel, S., and Tracey, J. (2016). Selection criteria for low resource language programs. In Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Correia, G. M., Niculae, V., and Martins, A. F. T. (2019). Adaptively sparse transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2174–2184, Hong Kong, China. Association for Computational Linguistics.

Craswell, N., Mitra, B., Yilmaz, E., Campos, D., and Voorhees, E. M. (2020). Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Dai, Z. and Callan, J. (2019). Deeper text understanding for ir with contextual neural language modeling. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 985–988.

Davis, M. W. and Dunning, T. E. (1995). A trec evaluation of query translation methods for multi-lingual text retrieval. In *Fourth Text Retrieval Conference*, volume 483.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser, L. (2019). Universal transformers. In *International Conference on Learning Representations*.

Dehghani, M., Zamani, H., Severyn, A., Kamps, J., and Croft, W. B. (2017). Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74.

Déjean, H., Gaussier, E., and Sadat, F. (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *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. Association for Computational Linguistics.

El-Kishky, A., Chaudhary, V., Guzmán, F., and Koehn, P. (2020a). CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969.

El-Kishky, A., Koehn, P., and Schwenk, H. (2020b). Searching the web for cross-lingual parallel data. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2417–2420, New York, NY, USA. Association for Computing Machinery.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2022). Language-agnostic BERT sentence embedding. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Ferreira, T. C., van der Lee, C., Van Miltenburg, E., and Krahmer, E. (2019). Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. *arXiv preprint arXiv:1908.09022*.

Gao, J. and Nie, J.-Y. (2006). A study of statistical models for query translation: Finding a good unit of translation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 194–201, New York, NY, USA. Association for Computing Machinery.

Gao, J., Zhou, M., Nie, J.-Y., He, H., and Chen, W. (2002). Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 183–190.

Gao, L. and Callan, J. (2022). Unsupervised corpus aware language model pre-training for dense passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Garera, N., Callison-Burch, C., and Yarowsky, D. (2009). Improving translation lexicon induction from monolingual corpora via dependency contexts and part-of-speech equivalences. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, page 129–137, USA. Association for Computational Linguistics.

Gaussier, E., Renders, J.-M., Matveeva, I., Goutte, C., and Déjean, H. (2004). A geometric view on bilingual lexicon extraction from comparable corpora. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 526–es, USA. Association for Computational Linguistics.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Grishman, R., Hirschman, L., and Nhan, N. T. (1986). Discovery procedures for sublanguage selectional patterns: Initial experiments. *Computational Linguistics*, 12(3):205–215.

Gritta, M. and Iacobacci, I. (2021). XeroAlign: Zero-shot cross-lingual transformer alignment. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 371–381, Online. Association for Computational Linguistics.

Guo, J., Fan, Y., Ai, Q., and Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 55–64, New York, NY, USA. Association for Computing Machinery.

Guo, M., Dai, Z., Vrandečić, D., and Al-Rfou, R. (2020). Wiki-40B: Multilingual language model dataset. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

Hazem, A. and Morin, E. (2014). Improving bilingual lexicon extraction from comparable corpora using window-based and syntax-based models. In *Computational Linguistics and Intelligent Text Processing: 15th International Conference, CICLing 2014, Kathmandu, Nepal, April 6-12, 2014, Proceedings, Part II 15*, pages 310–323. Springer.

Heffernan, K., Çelebi, O., and Schwenk, H. (2022). Bitext mining using distilled sentence representations for low-resource languages. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Huang, Z., Bonab, H., Sarwar, S. M., Rahimi, R., and Allan, J. (2021). Mixed attention transformer for leveraging word-level knowledge to neural cross-lingual information retrieval. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 760–770.

Huang, Z., Yu, P., and Allan, J. (2023a). Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1048–1056, New York, NY, USA. Association for Computing Machinery.

Huang, Z., Zeng, H., Zamani, H., and Allan, J. (2023b). Soft prompt decoding for multilingual dense retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1208–1218, New York, NY, USA. Association for Computing Machinery.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2021). Unsupervised dense information retrieval with contrastive learning.

Izacard, G., Caron, M., Hosseini, L., Riedel, S., Bojanowski, P., Joulin, A., and Grave, E. (2022). Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., and Carreira, J. (2021). Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Johnson, J., Douze, M., and Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kamholz, D., Pool, J., and Colowick, S. M. (2014). Panlex: Building a resource for panlingual lexical translation. In *LREC*, pages 3145–3150.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Khattab, O. and Zaharia, M. (2020). Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Lawrie, D., MacAvaney, S., Mayfield, J., McNamee, P., Oard, D. W., Soldaini, L., and Yang, E. (2024). Overview of the trec 2023 neuclir track. *arXiv preprint arXiv:2404.08071*.

Lawrie, D., Mayfield, J., Oard, D. W., and Yang, E. (2022). Hc4: A new suite of test collections for ad hoc clir. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, pages 351–366. Springer.

Lawrie, D., Yang, E., Oard, D. W., and Mayfield, J. (2023). Neural approaches to multilingual information retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, page 521–536, Berlin, Heidelberg. Springer-Verlag.

Le Calvé, A. and Savoy, J. (2000). Database merging strategy based on logistic regression. *Information Processing & Management*, 36(3):341–359.

Lewis, P., Oğuz, B., Rinott, R., Riedel, S., and Schwenk, H. (2019). Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Li, B. and Cheng, P. (2018). Learning neural representation for CLIR with adversarial framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1861–1870, Brussels, Belgium. Association for Computational Linguistics.

Li, Y., Franz, M., Sultan, M. A., Iyer, B., Lee, Y.-S., and Sil, A. (2022). Learning cross-lingual IR from an English retriever. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4428–4436, Seattle, United States. Association for Computational Linguistics.

Litschko, R., Glavaš, G., Vulic, I., and Dietz, L. (2019). Evaluating resource-lean cross-lingual embedding models in unsupervised retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1109–1112, New York, NY, USA. Association for Computing Machinery.

Litschko, R., Vulić, I., and Glavaš, G. (2022a). Parameter-efficient neural reranking for cross-lingual and multilingual retrieval. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1071–1082, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Litschko, R., Vulić, I., Ponzetto, S. P., and Glavaš, G. (2022b). On cross-lingual retrieval with multilingual text encoders. *Information Retrieval Journal*, 25(2):149–183.

Liu, Y., Jin, R., and Chai, J. Y. (2005). A maximum coherence model for dictionary-based cross-language information retrieval. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '05, page 536–543, New York, NY, USA. Association for Computing Machinery.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

MacAvaney, S., Yates, A., Cohan, A., and Goharian, N. (2019). Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1101–1104.

Maeda, A., Sadat, F., Yoshikawa, M., and Uemura, S. (2000). Query term disambiguation for web cross-language information retrieval using a search engine. In *Proceedings of the Fifth International Workshop on on Information Retrieval with Asian Languages*, IRAL '00, page 25–32, New York, NY, USA. Association for Computing Machinery.

Manmatha, R., Rath, T., and Feng, F. (2001). Modeling score distributions for combining the outputs of search engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 267–275, New York, NY, USA. Association for Computing Machinery.

McDonnell, J. R., Reynolds, R. G., and Fogel, D. B. (1995). *Query Translation Using Evolutionary Programming for Multi-Lingual Information Retrieval*, pages 175–185.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. In Bengio, Y. and LeCun, Y., editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Nair, S., Yang, E., Lawrie, D., Duh, K., McNamee, P., Murray, K., Mayfield, J., and Oard, D. W. (2022). Transfer learning approaches for building cross-language dense retrieval models. In *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, pages 382–396. Springer.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

Nie, J.-Y. (2022). *Cross-language information retrieval*. Springer Nature.

Nogueira, R. and Cho, K. (2019). Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Nooralahzadeh, F., Bekoulis, G., Bjerva, J., and Augenstein, I. (2020). Zero-shot cross-lingual transfer with meta learning. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4547–4562, Online. Association for Computational Linguistics.

Oard, D. W. and Diekema, A. R. (1998). Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, 33:223–56.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Ott, M., Edunov, S., Grangier, D., and Auli, M. (2018). Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global vectors for word representation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Peters, C. (2005). What happened in clef 2004? In Peters, C., Clough, P., Gonzalo, J., Jones, G. J. F., Kluck, M., and Magnini, B., editors, *Multilingual Information Access for Text, Speech and Images*, pages 1–9, Berlin, Heidelberg. Springer Berlin Heidelberg.

Peters, C. (2006). What happened in clef 2005. In Peters, C., Gey, F. C., Gonzalo, J., Müller, H., Jones, G. J. F., Kluck, M., Magnini, B., and de Rijke, M., editors, *Accessing Multilingual Information Repositories*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

Peters, C. (2007). What happened in clef 2006. In Peters, C., Clough, P., Gey, F. C., Karlgren, J., Magnini, B., Oard, D. W., de Rijke, M., and Stempfhuber, M., editors, *Evaluation of Multilingual and Multi-modal Information Retrieval*, pages 1–10, Berlin, Heidelberg. Springer Berlin Heidelberg.

Peters, C. (2008). What happened in clef 2007. In Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D. W., Peñas, A., Petras, V., and Santos, D., editors, *Advances in Multilingual and Multimodal Information Retrieval*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Peters, C. (2009). What happened in clef 2008. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G. J. F., Kurimo, M., Mandl, T., Peñas, A., and Petras, V., editors, *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 1–14, Berlin, Heidelberg. Springer Berlin Heidelberg.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Pires, T., Schlinger, E., and Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Qiao, Y., Xiong, C., Liu, Z., and Liu, Z. (2019). Understanding the behaviors of bert in ranking.

Qin, L., Ni, M., Zhang, Y., and Che, W. (2021). Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Rahimi, R., Montazeralghaem, A., and Shakery, A. (2020). An axiomatic approach to corpus-based cross-language information retrieval. *Information Retrieval Journal*, 23:191–215.

Rahimi, R., Shakery, A., and King, I. (2015). Multilingual information retrieval in the language modeling framework. *Information Retrieval Journal*, 18:246–281.

Rahimi, R., Shakery, A., and King, I. (2016). Extracting translations from comparable corpora for cross-language information retrieval using the language modeling framework. *Inf. Process. Manage.*, 52(2):299–318.

Ravfogel, S., Goldberg, Y., and Cotterell, R. (2023). Log-linear guardedness and its implications. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, Toronto, Canada. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Roy, U., Constant, N., Al-Rfou, R., Barua, A., Phillips, A., and Yang, Y. (2020). LAReQA: Language-agnostic answer retrieval from a multilingual pool. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5919–5930, Online. Association for Computational Linguistics.

Ruder, S., Clark, J., Gutkin, A., Kale, M., Ma, M., Nicosia, M., Rijhwani, S., Riley, P., Sarr, J.-M., Wang, X., Wieting, J., Gupta, N., Katanova, A., Kirov, C., Dickinson, D., Roark, B., Samanta, B., Tao, C., Adelani, D., Axelrod, V., Caswell, I., Cherry, C., Garrette, D., Ingle, R., Johnson, M., Panteleev, D., and Talukdar, P. (2023). XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Sadat, F., Yoshikawa, M., and Uemura, S. (2003). Enhancing cross-language information retrieval by an automatic acquisition of bilingual terminology from comparable corpora. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval*, pages 397–398.

Salton, G. (1970). Automatic processing of foreign language documents. *Journal of the American Society for Information Science*, 21(3):187–194.

Salton, G. (1973). Experiments in multi-lingual information retrieval. *Information Processing Letters*, 2(1):6–11.

Sarwar, S. M., Bonab, H., and Allan, J. (2019). A multi-task architecture on relevance-based neural query translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6339–6344, Florence, Italy. Association for Computational Linguistics.

Sasaki, S., Sun, S., Schamoni, S., Duh, K., and Inui, K. (2018). Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463.

Savoy, J. (2003). Report on clef 2002 experiments: Combining multiple sources of evidence. In *Advances in Cross-Language Information Retrieval: Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002 Rome, Italy, September 19–20, 2002 Revised Papers 3*, pages 66–90. Springer.

Savoy, J. and Berger, P.-Y. (2005). Selection and merging strategies for multilingual information retrieval. In *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers 5*, pages 27–37. Springer.

Sorg, P. and Cimiano, P. (2012). Exploiting wikipedia for cross-lingual and multilingual information retrieval. *Data Knowl. Eng.*, 74:26–45.

Su, Y., Wang, X., Qin, Y., Chan, C.-M., Lin, Y., Wang, H., Wen, K., Liu, Z., Li, P., Li, J., et al. (2022). On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969.

Sun, S. and Duh, K. (2020). CLIRMatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170, Online. Association for Computational Linguistics.

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Tenney, I., Das, D., and Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Thakur, N., Ni, J., Ábrego$^{\diamondsuit}$, G. H., Wieting, J., Lin, J., and Cer, D. (2023). Leveraging llms for synthesizing training data across many languages in multilingual dense retrieval. *CoRR*, abs/2311.05800.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Tiyajamorn, N., Kajiwara, T., Arase, Y., and Onizuka, M. (2021). Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Vu, T., Lester, B., Constant, N., Al-Rfou, R., and Cer, D. M. (2021). Spot: Better frozen model adaptation through soft prompt transfer. In *Annual Meeting of the Association for Computational Linguistics*.

Vulić, I. and Moens, M.-F. (2015). Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 363–372, New York, NY, USA. Association for Computing Machinery.

Wang, Z., K, K., Mayhew, S., and Roth, D. (2020). Extending multilingual BERT to low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Wang, Z., Panda, R., Karlinsky, L., Feris, R., Sun, H., and Kim, Y. (2023). Multitask prompt tuning enables parameter-efficient transfer learning. In *The Eleventh International Conference on Learning Representations*.

Wu, L., Wu, S., Zhang, X., Xiong, D., Chen, S., Zhuang, Z., and Feng, Z. (2022). Learning disentangled semantic representations for zero-shot cross-lingual transfer in multilingual machine reading comprehension. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 991–1000, Dublin, Ireland. Association for Computational Linguistics.

Wu, S. and Dredze, M. (2019). Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Wu, S. and Dredze, M. (2020). Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Xie, Y., Wang, X., Wang, R., and Zha, H. (2020). A fast proximal point method for computing exact wasserstein distance. In Adams, R. P. and Gogate, V., editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*. PMLR.

Xie, Z., Zhao, H., Yu, T., and Li, S. (2022). Discovering low-rank subspaces for language-agnostic multilingual representations. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5617–5633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiong, C., Dai, Z., Callan, J., Liu, Z., and Power, R. (2017). End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 55–64, New York, NY, USA. Association for Computing Machinery.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P. N., Ahmed, J., and Overwijk, A. (2021). Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Xu, J., Weischedel, R., and Nguyen, C. (2001). Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 105–110.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Hernandez Abrego, G., Yuan, S., Tar, C., Sung, Y.-h., Strope, B., and Kurzweil, R. (2020). Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Yao, L., Yang, B., Zhang, H., Chen, B., and Luo, W. (2020a). Domain transfer based data augmentation for neural query translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4521–4533.

Yao, L., Yang, B., Zhang, H., Luo, W., and Chen, B. (2020b). Exploiting neural query translation into cross lingual information retrieval. *arXiv preprint arXiv:2010.13659*.

Yates, A., Nogueira, R., and Lin, J. (2021). Pretrained transformers for text ranking: BERT and beyond. In Kondrak, G., Bontcheva, K., and Gillick, D., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 1–4, Online. Association for Computational Linguistics.

Yu, P., Fei, H., and Li, P. (2021). Cross-lingual language model pretraining for retrieval. In *Proceedings of the Web Conference 2021*, WWW '21, page 1029–1039, New York, NY, USA. Association for Computing Machinery.

Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., and Ma, S. (2021). Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512.

Zhan, J., Mao, J., Liu, Y., Zhang, M., and Ma, S. (2020). An analysis of bert in document ranking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1941–1944, New York, NY, USA. Association for Computing Machinery.

Zhang, L., Karakos, D., Hartmann, W., Srivastava, M., Tarlin, L., Akodes, D., Gouda, S. K., Bathool, N., Zhao, L., Jiang, Z., Schwartz, R., and Makhoul, J. (2020). The 2019 BBN cross-lingual information retrieval system. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 44–51, Marseille, France. European Language Resources Association.

Zhang, L. and Zhao, X. (2020). An overview of cross-language information retrieval. In Sun, X., Wang, J., and Bertino, E., editors, *Artificial Intelligence and Security*, pages 26–37, Cham. Springer International Publishing.

Zhang, X., Ma, X., Shi, P., and Lin, J. (2021). Mr. TyDi: A multi-lingual benchmark for dense retrieval. In Ataman, D., Birch, A., Conneau, A., Firat, O., Ruder, S., and Sahin, G. G., editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., and Lin, J. (2023). MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Zhao, Y. and Bethard, S. (2020). How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.