Multilingual Information Retrieval (MLIR)

601.764

2/14/2023

Much of the lecture comes from Lawrie et al., 2023

“Neural Approaches to Multilingual Information Retrieval”

CLIR → MLIR

- What's the difference?
- What's the definition?

MLIR

- Monolingual Retrieval in Multiple Languages
- Query to construct one ranked list where each document is in one of several languages
  - Cross-Language Evaluation Forum (CLEF)
- Mixed-Language Queries
- Mixed-Language Documents

MLIR

Querying Across Languages: A Dictionary-Based Approach to Multilingual Information Retrieval

David A. Hull **Gregory Grefenstette**

1996

2 Defining Multilingual Information Retrieval

There is no common currently accepted definition for multilingual information retrieval (MLIR). The term has been used in the past to cover a broad range of different approaches to information retrieval (IR) using one or more languages. In this section, we will present a number of different descriptions of the multilingual IR task that have been used by previous authors and outline our own approach to the problem. Five different definitions for MLIR are outlined below.

(1) IR in any language other than English.

(2) IR on a parallel document collection or on a multilingual document collection where the search space is restricted to the query language.

(3) IR on a monolingual document collection which can be queried in multiple languages.

(4) IR on a multilingual document collection, where queries can retrieve documents in multiple languages.

(5) IR on multilingual documents, i.e. more than one language can be present in the individual documents

Datasets

CLEF

Table 1. Dataset statistics of CLEF 2001, 2002, and 2003. CLEF 2001 and 2002 share the document collection but have different queries. Numbers in parentheses are the number of topics in each query set. We report the number of documents judged relevant over all the topics in a particular year.

Query Set	English		German		Spanish		French		Italian		Total	
	# Rel.	# Docs	# Rel.	# Docs	# Rel.	# Docs	# Rel.	# Docs	# Rel.	# Docs	# Rel.	# Docs
2001 (50)	856	113,005	2,130	225,371	2,694	215,738	1,212	87,191	1,246	108,578	8,138	749,883
2002 (50)	821		1,938		2,854		1,383		1,072		8,068	
2003 (60)	1,006	169,477	1,825	294,809	2,367	454,045	946	129,806	–	–	6,144	1,048,137

MS MARCO

Follow @MSMarcoAI

Starting with a paper released at NIPS 2016, MS MARCO is a collection of datasets focused on deep learning in search.

The first dataset was a question answering dataset featuring 100,000 real Bing questions and a human generated answer. Since then we released a 1,000,000 question dataset, a natural langauge generation dataset, a passage ranking dataset, keyphrase extraction dataset, crawling dataset, and a conversational search.

2022

mMARCO

mMARCO: A Multilingual Version of the MS MARCO Passage Ranking Dataset

Luiz Bonifacio
Univ. of Campinas
NeuralMind

Vitor Jeronymo
Univ. of Campinas
NeuralMind

Hugo Queiroz Abonizio
NeuralMind

Israel Campiotti
NeuralMind

Marzieh Fadaee
Zeta Alpha

Roberto Lotufo
Univ. of Campinas
NeuralMind

Rodrigo Nogueira
Univ. of Campinas
Univ. of Waterloo
NeuralMind

13 Languages

"Helsinki" models From Hugging Face

Google Translate API

2021

Mr. TyDi

Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin

David R. Cheriton School of Computer Science
University of Waterloo

11 Languages

Monolingual

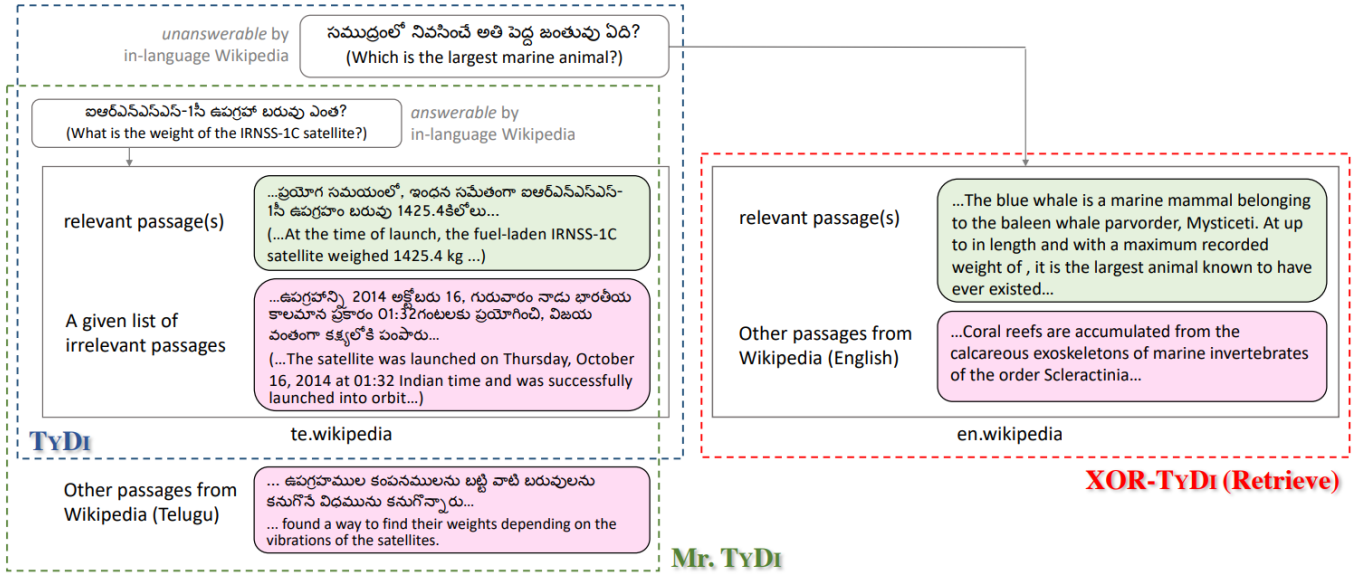


Figure 1: Comparison between TyDi, XOR-TyDi, and Mr. TyDi with an example in Telugu. The green blocks indicate relevant passages and the red blocks indicate non-relevant passages.

NeuCLIR

- 2022
- 2023 (On-Going)
- Russian, Chinese, Farsi

Evaluation

- MAP
- nDCG
- P@10

Result Pooling

- Manual Examination of all Documents is infeasible
- Top-Ranked Documents from Multiple IR systems (k=100 or 1,000)
- Multiple Assessors
- Can be Boolean (or not)
- Cohen's Kappa
- Mate Retrieval in Multilingual (Bitexts/Translations Used)

5 Broad Approaches

1997

1.) Embeddings

Lower Rank Approximation using SVD

Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing

Bob Rehder
Dept. of Psychology
Inst. of Cognitive Science
U. of Colorado, Boulder
Boulder, CO 80309
rehder@psych.colorado.edu

Michael L. Littman
Dept. of Computer Sci.

Duke University
Durham, NC 27708
mlittman@cs.duke.edu

Susan Dumais
Microsoft Research

One Microsoft Way
Redmond WA, 98052
sdumais@microsoft.com

Thomas K. Landauer
Dept. of Psychology
Inst. of Cognitive Science
U. of Colorado, Boulder
Boulder, CO 80309
landauer@psych.colorado.edu

Log-entropy weighting

Query	Tempest	Romeo	Hamlet	King Lear	Midsummer	Macbeth
0	0	1	0	1	0	0
0	1	0	0	0	0	0
1	1	0	0	0	0	0
0	0	0	0	0	0	0

Cosine Similarity

1.) Embeddings

- Rehder et al. 1997
- Gabrilovich and Markovitch 2007
- Sorg and Cimiano 2012
- ….
- Modern Language Model Approaches

2.) Query in All Languages

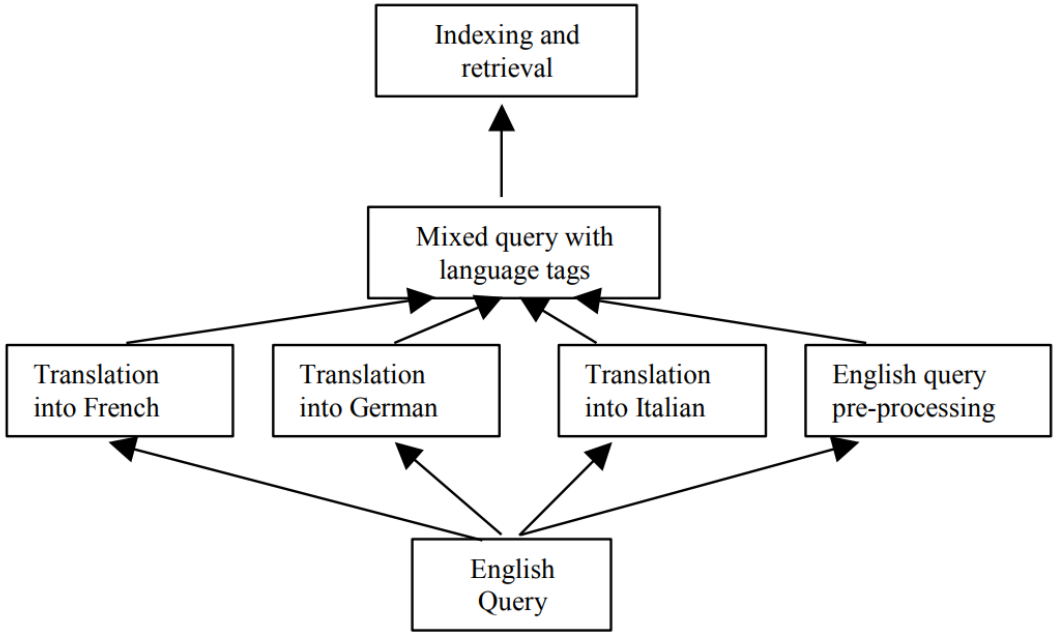
Problems with Scale of # Langs

A Multilingual Approach to Multilingual Information Retrieval

2002

Jian-Yun Nie, Fuman Jin

Laboratoire RALI
Département d'Informatique et Recherche opérationnelle,
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Québec, H3C 3J7 Canada
{nie, jinf}@iro.umontreal.ca



3.) Translate Index

4.) Pivot Languages

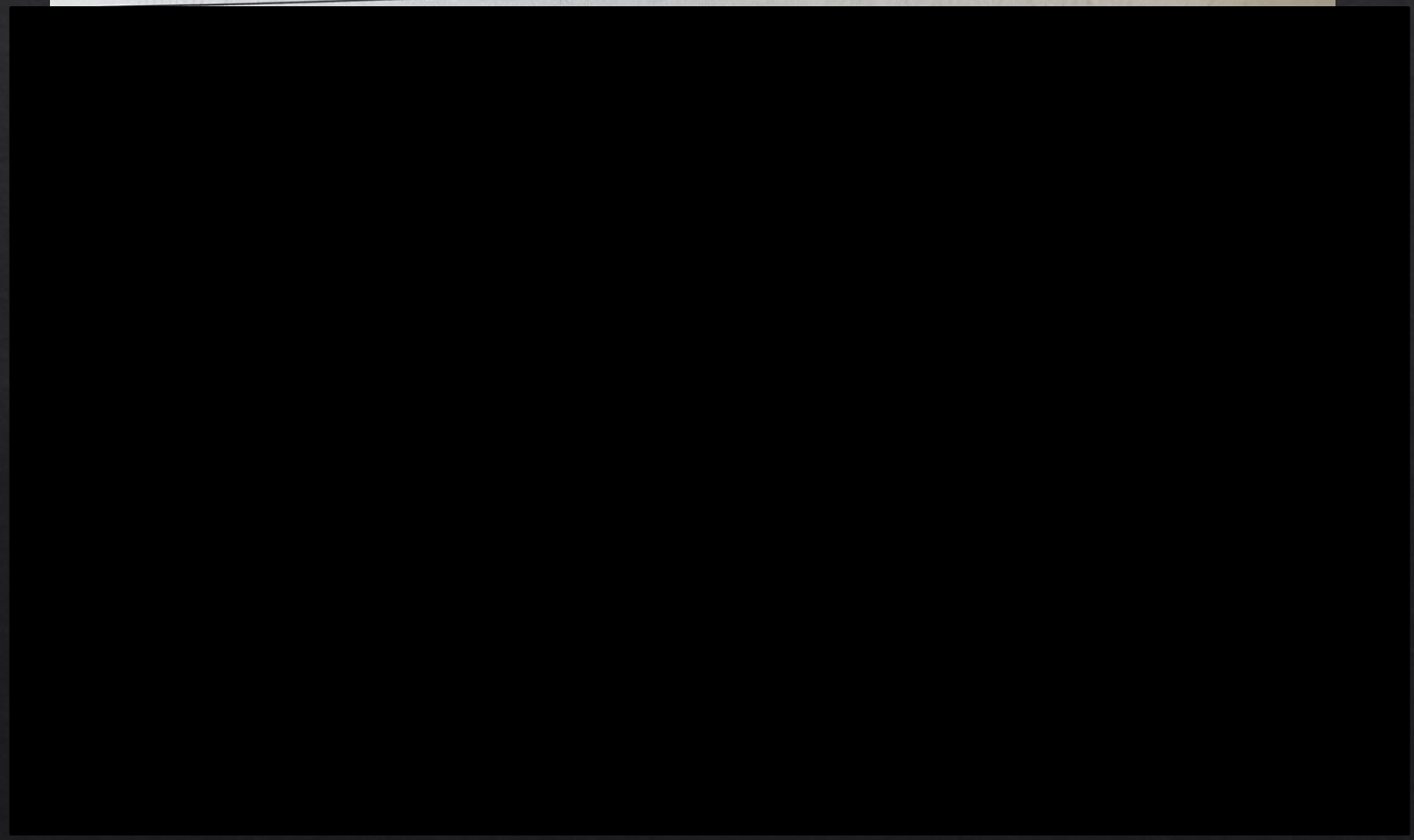
- Translate Queries into Subset of Languages
- Translate Documents into Subset of Languages
- Extreme Case → Only 1 Language

5.) Merge Ranked Lists

- Most Widely Studied
- Monolingual or Bilingual Retrieval to create a list for each document language
- Similar to Collection Sharding (efficiency method)
- … but requires normalization prior to late fusion
- Normalization is challenging!

Normalization

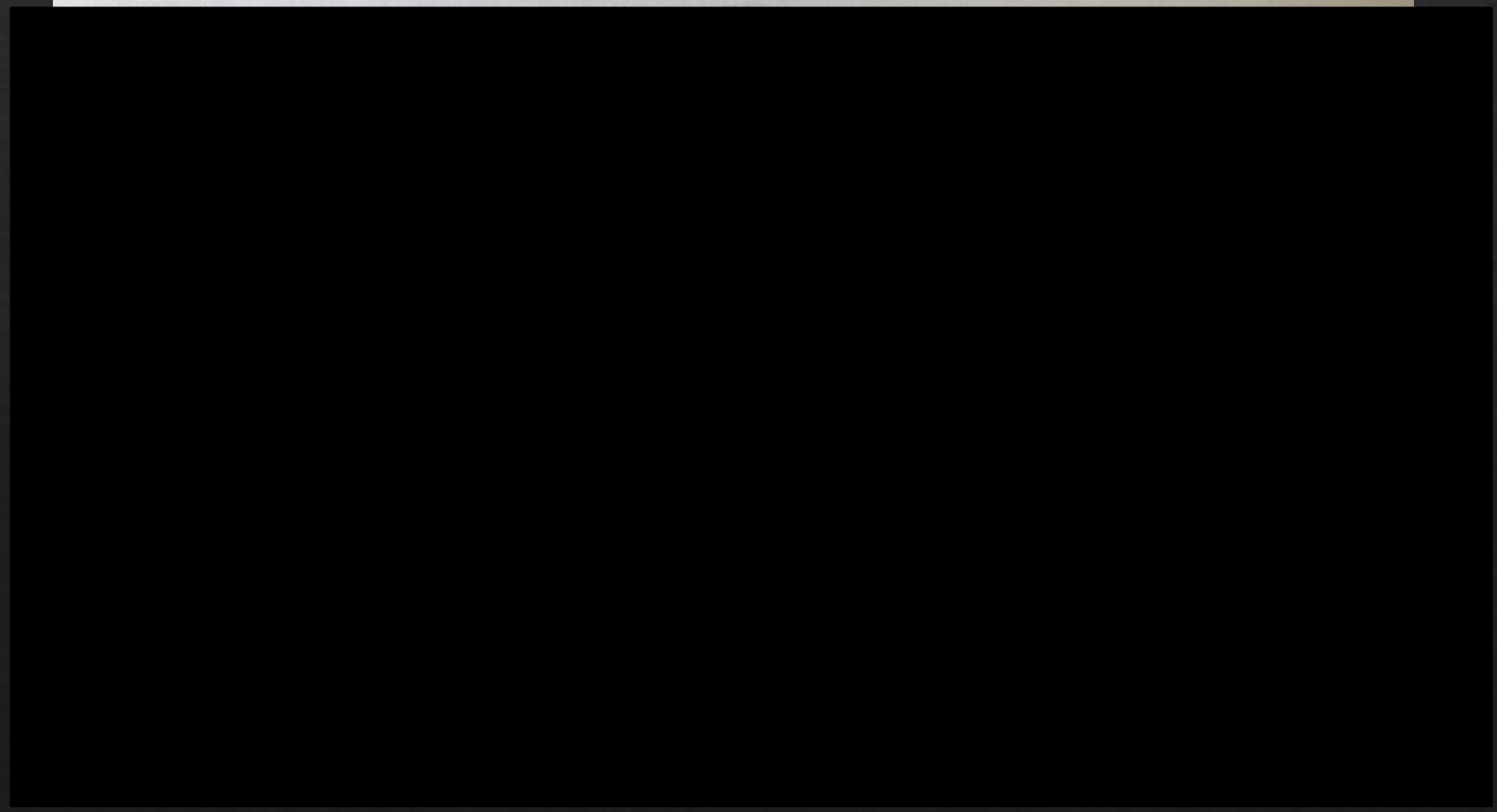
Algorithm 11–2 Aggregation of multiple rankings $r_1, \ldots, r_n$ based on Z-score normalization. For ranking r, $r[i]$ defines the score at rank position i, $\text{score}_r(d)$ defines the score of document d. MIN, MEAN, and STD-DEVIATION are defined on the set of score values of ranking r



Normalization

Algorithm 11–2 Aggregation of multiple rankings $r_1, \ldots, r_n$ based on Z-score normalization. For ranking r, $r[i]$ defines the score at rank position i, $\text{score}_r(d)$ defines the score of document d. MIN, MEAN, and STD-DEVIATION are defined on the set of score values of ranking r

$R \leftarrow \{r_1, \ldots, r_n\}$
for all $r \in R$ **do** // *Normalization*



Normalization

Algorithm 11–2 Aggregation of multiple rankings $r_1, \ldots, r_n$ based on Z-score normalization. For ranking r, $r[i]$ defines the score at rank position i, $\text{score}_r(d)$ defines the score of document d. MIN, MEAN, and STD-DEVIATION are defined on the set of score values of ranking r

$R \leftarrow \{r_1, \ldots, r_n\}$
for all $r \in R$ **do** // *Normalization*
  $\mu \leftarrow \text{MEAN}(r)$
  $\sigma \leftarrow \text{STD-DEVIATION}(r)$
  $\delta \leftarrow \frac{\mu - \text{MIN}(r)}{\sigma}$

Normalization

Algorithm 11–2 Aggregation of multiple rankings $r_1, \ldots, r_n$ based on Z-score normalization. For ranking r, $r[i]$ defines the score at rank position i, $\text{score}_r(d)$ defines the score of document d. MIN, MEAN, and STD-DEVIATION are defined on the set of score values of ranking r

$R \leftarrow \{r_1, \ldots, r_n\}$
for all $r \in R$ **do** // *Normalization*
  $\mu \leftarrow \text{MEAN}(r)$
  $\sigma \leftarrow \text{STD-DEVIATION}(r)$
  $\delta \leftarrow \frac{\mu - \text{MIN}(r)}{\sigma}$
  for $i = 1..|r|$ **do**
    $r(i) \leftarrow \frac{r(i) - \mu}{\sigma} + \delta$
  end for
end for

[redacted]

Normalization

Algorithm 11–2 Aggregation of multiple rankings $r_1, \ldots, r_n$ based on Z-score normalization. For ranking r, $r[i]$ defines the score at rank position i, $\text{score}_r(d)$ defines the score of document d. MIN, MEAN, and STD-DEVIATION are defined on the set of score values of ranking r

$R \leftarrow \{r_1, \ldots, r_n\}$
for all $r \in R$ **do** // *Normalization*
  $\mu \leftarrow \text{MEAN}(r)$
  $\sigma \leftarrow \text{STD-DEVIATION}(r)$
  $\delta \leftarrow \frac{\mu - \text{MIN}(r)}{\sigma}$
  for $i = 1..|r|$ **do**
    $r(i) \leftarrow \frac{r(i)-\mu}{\sigma} + \delta$
  end for
end for

$r_c \leftarrow \{\}$
for all $d \in D$ **do** // *Aggregation*
  $s \leftarrow 0$
  for all $r \in R$ **do**
    $s \leftarrow s + \text{score}_r(d)$

Normalization

Algorithm 11–2 Aggregation of multiple rankings $r_1, \ldots, r_n$ based on Z-score normalization. For ranking r, $r[i]$ defines the score at rank position i, $\text{score}_r(d)$ defines the score of document d. MIN, MEAN, and STD-DEVIATION are defined on the set of score values of ranking r

$R \leftarrow \{r_1, \ldots, r_n\}$
for all $r \in R$ **do** // *Normalization*
  $\mu \leftarrow \text{MEAN}(r)$
  $\sigma \leftarrow \text{STD-DEVIATION}(r)$
  $\delta \leftarrow \frac{\mu - \text{MIN}(r)}{\sigma}$
  for $i = 1..|r|$ **do**
    $r(i) \leftarrow \frac{r(i)-\mu}{\sigma} + \delta$
  end for
end for

$r_c \leftarrow \{\}$
for all $d \in D$ **do** // *Aggregation*
  $s \leftarrow 0$
  for all $r \in R$ **do**
    $s \leftarrow s + \text{score}_r(d)$
  end for
  $\text{score}_{r_c}(d) \leftarrow s$
end for

[redacted]

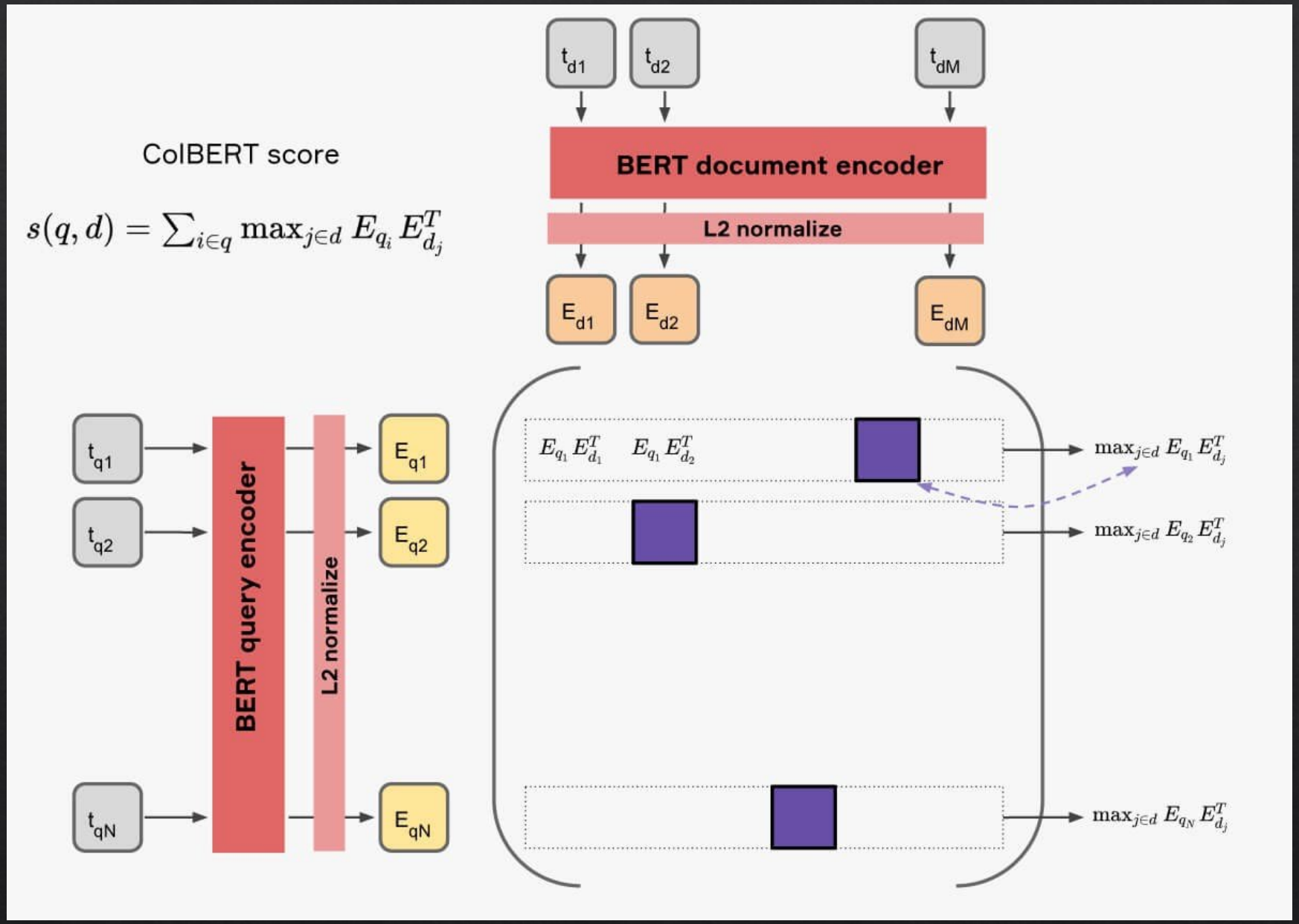
Normalization

Algorithm 11–2 Aggregation of multiple rankings $r_1, \ldots, r_n$ based on Z-score normalization. For ranking r, $r[i]$ defines the score at rank position i, $\text{score}_r(d)$ defines the score of document d. MIN, MEAN, and STD-DEVIATION are defined on the set of score values of ranking r

$R \leftarrow \{r_1, \ldots, r_n\}$
for all $r \in R$ **do** // *Normalization*
  $\mu \leftarrow \text{MEAN}(r)$
  $\sigma \leftarrow \text{STD-DEVIATION}(r)$
  $\delta \leftarrow \frac{\mu - \text{MIN}(r)}{\sigma}$
  for $i = 1..|r|$ **do**
    $r(i) \leftarrow \frac{r(i)-\mu}{\sigma} + \delta$
  end for
end for

$r_c \leftarrow \{\}$
for all $d \in D$ **do** // *Aggregation*
  $s \leftarrow 0$
  for all $r \in R$ **do**
    $s \leftarrow s + \text{score}_r(d)$
  end for
  $\text{score}_{r_c}(d) \leftarrow s$
end for
$r_c \leftarrow \text{DESCENDING-SORT}(r_c)$
return r_c

But we are in the neural world…



https://medium.com/@varun030403/colbert-a-complete-guide-1552468335ae

mBERT (Multilingual BERT)

NeuCLIR

Devlin et al., 2018

Multilingual Translate Training (MTT)

- Translate monolingual training data to target languages
- MTT-S (Batches contain Single-Language)
- MTT-M (Multilingual Batches)
- … and you can have zero-shot

CLEF ColBERT-X MTT

	MAP			P@10		
	2001	2002	2003	2001	2002	2003
MTT-M	**0.462†**	**0.462†**	**0.461†**	**0.704**	**0.752**	**0.653**
MTT-S	0.422	0.405	0.433	0.696	0.702	0.649

CLEF MAP

Query Set	ITD	MAP MULM	BM25	ColBERT-X MTT-M	ET
Title Queries					
2001	✓	–	***0.398***	0.377	0.391
	✗	0.349	–	*0.360*	0.322
2002	✓	–	0.337	0.367	***0.389***
	✗	0.276	–	*0.352*	0.333
2003	✓	–	***0.349***	0.337	***0.349***
	✗	0.305	–	*0.332*	0.290
All	✓	–	0.361	0.359	***0.375***
	✗	0.310	–	*0.347*	0.314†

Speed

Table 6. ColBERT-X GPU hours for translating and indexing. BM25 does not use GPU.

Model	ITD	CLEF2001-2002 Translation	Index	Total	CLEF2003 Translation	Index	Total
BM25	✓	55.0	–	55.0	68.6	–	68.6
ET	✓	[illegible]	[illegible]	[illegible]	[illegible]	12.3	80.9
	✗	–	9.9	9.9	–	12.4	12.4
MT-5	✓	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	87.6
	✗	–	16.7	16.7	–	21.9	21.9
MT-N	✓	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	✗	–	13.1	13.1	–	19.3	19.3

Translating the whole corpus is slow …

… and maybe not worth it

Language Bias

- Exhibits Bias towards language trained on
- 😮
- Document Scores not drawn from the same distribution
- All Western European and Latin Script