Teaching a New Dog Old Tricks:
Resurrecting Multilingual Retrieval Using Zero-shot Learning

Sean MacAvaney, Luca Soldaini, Nazli Goharian

sean@ir.cs.georgetown.edu · https://macavaney.us · @macavaney

To appear at ECIR 2020 (short paper)

arxiv.org/abs/1912.13080

In IR we have a wide variety of datasets for training/evaluation of ad-hoc search.

“Deep” Datasets

(with few queries but many judgments)

E.g., TREC Robust

“Wide” Datasets

(with many queries but few judgments)

E.g., MS-MARCO

Domain Variety

News Article Search
e.g., TREC News Track

Web Document Search
e.g., TREC WebTrack / ClueWeb

Complex Answers
e.g., TREC CAR

Etc.

Well, at least we do in English…

The amount and diversity of non-English IR resources is low.

Language	Speakers *	Wide Datasets # queries	Deep datasets # queries	Domains	Public Ad-hoc Benchmarks
English	~1.1B	millions	thousands	News, Web, QA, Medical, …	MS-MARCO, TREC, …
Chinese Mandarin	~1.1B	~500k	300	News, Web	Sogou-QCL, TREC, NTCIR
Hindi	~620M	0	200	News	FIRE
Spanish	~530M	0	206	News	TREC, CLEF
French	~280M	0	333	News	CLEF
Arabic	~270M	0	75	News	TREC
Bengali	~270M	0	200	News	FIRE
Russian	~260M	0	62	News	CLEF
Portuguese	~230M	0	146	News	CLEF
Indonesian	~200M	0	0	-	-

* 2019 SIL Ethnologue

Recent trends in ad-hoc retrieval

- Neural models based on contextualized language models are very effective at document ranking.
  - E.g., Use classification mechanism of BERT to rank documents (i.e., “Vanilla BERT”)

- These models can be either trained on suitably large “deep” datasets or “wide” datasets.

- **Trained & evaluated on English datasets!**

Research Questions

- Can neural ranking methods be applied across languages to overcome scarcity of resources?

- Can we take advantage of contextualized language models trained on multiple languages for this task?

- Are old evaluation datasets in languages other than English still suitable for evaluation?

Experiment: Zero-shot Multilingual Retrieval

Step 1: Pre-Train contextualized language model on massive amount of multi-lingual data.

WIKIPEDIA
The Free Encyclopedia

English
5 994 000+ articles

日本語
1 185 000+ 記事

Deutsch
2 385 000+ Artikel

Español
1 571 000+ artículos

Русский
1 590 000+ статей

Français
2 171 000+ articles

Italiano
1 576 000+ voci

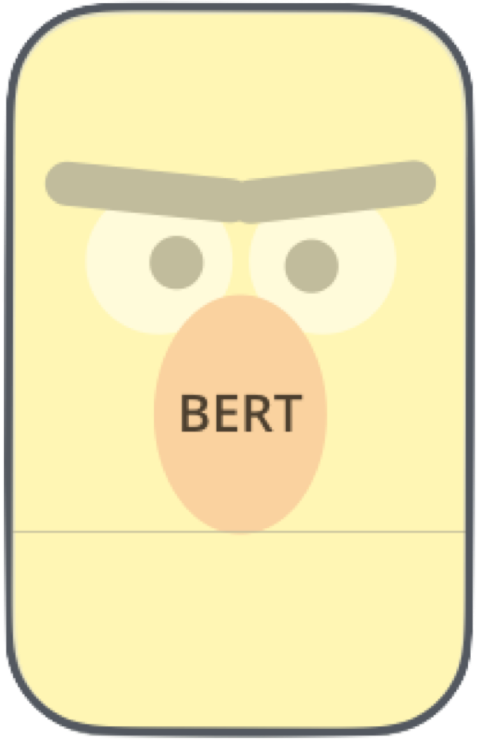
中文
1 090 000+ 條目

Português
1 018 000+ artigos

Polski
1 379 000+ haseł

- We use the pretrained `bert-base-multilingual-cased` model, which uses Wikipedia text from the largest 104 Wikipedias.
- Trained both with Masked Language Model (MLM) and Next Sentence Prediction (NSP) objectives.

Step 2: Fine-tune model to predict relevance based on English relevance pairs.

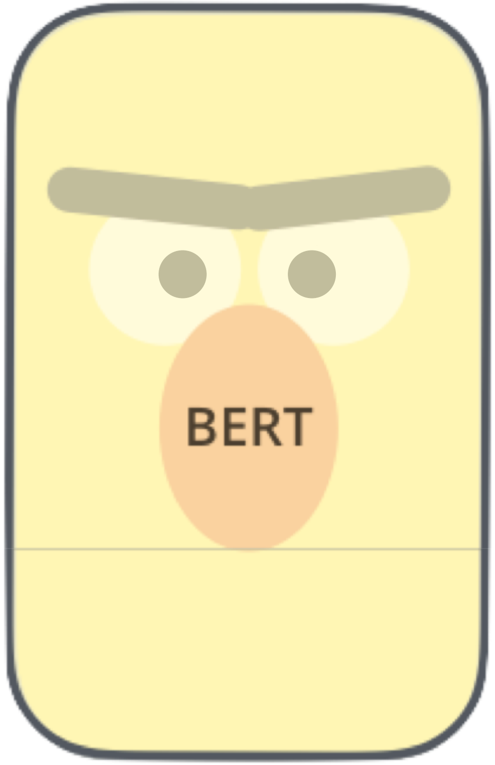


TREC Robust 2004

- Trained using pairwise loss, to learn to rank relevance documents higher than non-relevant documents.
- Optimizing score from language model’s [CLS] representation

Step 2: Fine-tune model to predict relevance based on English relevance pairs.

Query: **International Organized Crime**



Document 1: …This is the ninth and last time I will appear before you to fulfill my constitutional obligation of addressing you at the beginning of each legislative period to deliver a message on administrative matters…

Document 2: …Today, Colombian Prosecutor General Gustavo de Greiff said the U.S. Government is not interested in supplying evidence to condemn the chiefs of the drug trafficking mafia because it does not trust Colombian justice…

Step 3: Test by retrieving and re-ranking queries/documents for other languages.

Query: **Oposición Mexicana al TLC**

Document 1: …representantes sindicales de mexico informaron hoy que iniciaran un referendum en todo el pais…

Document 2: …la administracion de bill clinton presionara al gobierno mexicano para reabrir la negociacion del…

Document 3: …un estudio de heritage foundation indica que sera dificil la aprobacion del tratado de libre comercio…

Document 4: …roberto mercado la semana anterior se celebro en monterrey la conferencia internacional cuatro…

Document 5: …roberto mercado los negociadores de colombia , mexico y venezuela mantuvieron una reunion los …

- Initial retrieval using approach such as BM25.
- Using proper stemming/preprocessing per language.
- Using language model’s classification technique.
- Referred to as “Vanilla BERT”

Step 3: Test by Retrieving and re-ranking queries/documents for other languages.

Query: **Oposición Mexicana al TLC**

- Initial retrieval using approach such as BM25.
- Using proper stemming/preprocessing per language.
- Using language model’s classification technique.
- Referred to as “Vanilla BERT”

Experiment - Datasets

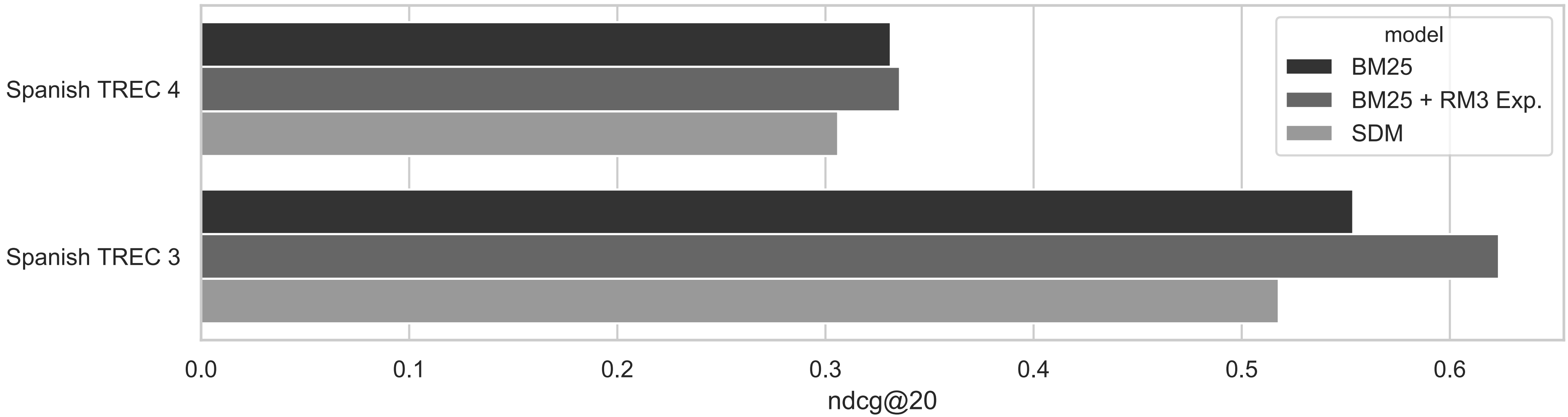
- We train on the following dataset

Dataset	# Topics	# Judgments	Document Collection
TREC Robust 2004 (English)	249	311k	528k news articles from TREC Disks 4+5

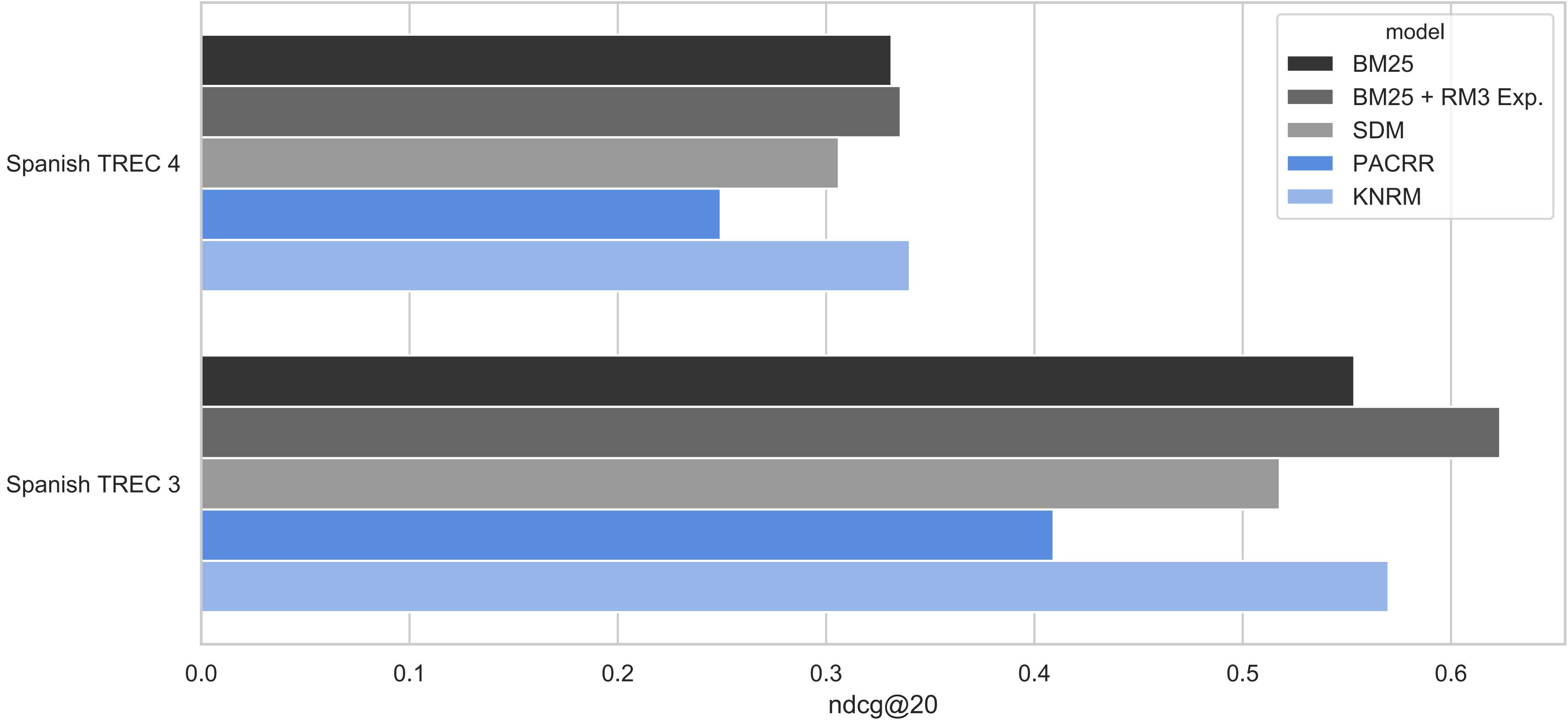
- We test on the following datasets

Dataset	# Topics	# Judgments	Document Collection
TREC Spanish 3	25	19k	58k news articles from LDC2000T51
TREC Spanish 4	25	13k	
TREC Mandarin 5	26	16k	130k news articles from LDC2000T52
TREC Mandarin 6	28	9k	
TREC Arabic 2001	25	23k	384k news articles from LDC2001T55
TREC Arabic 2002	50	38k	

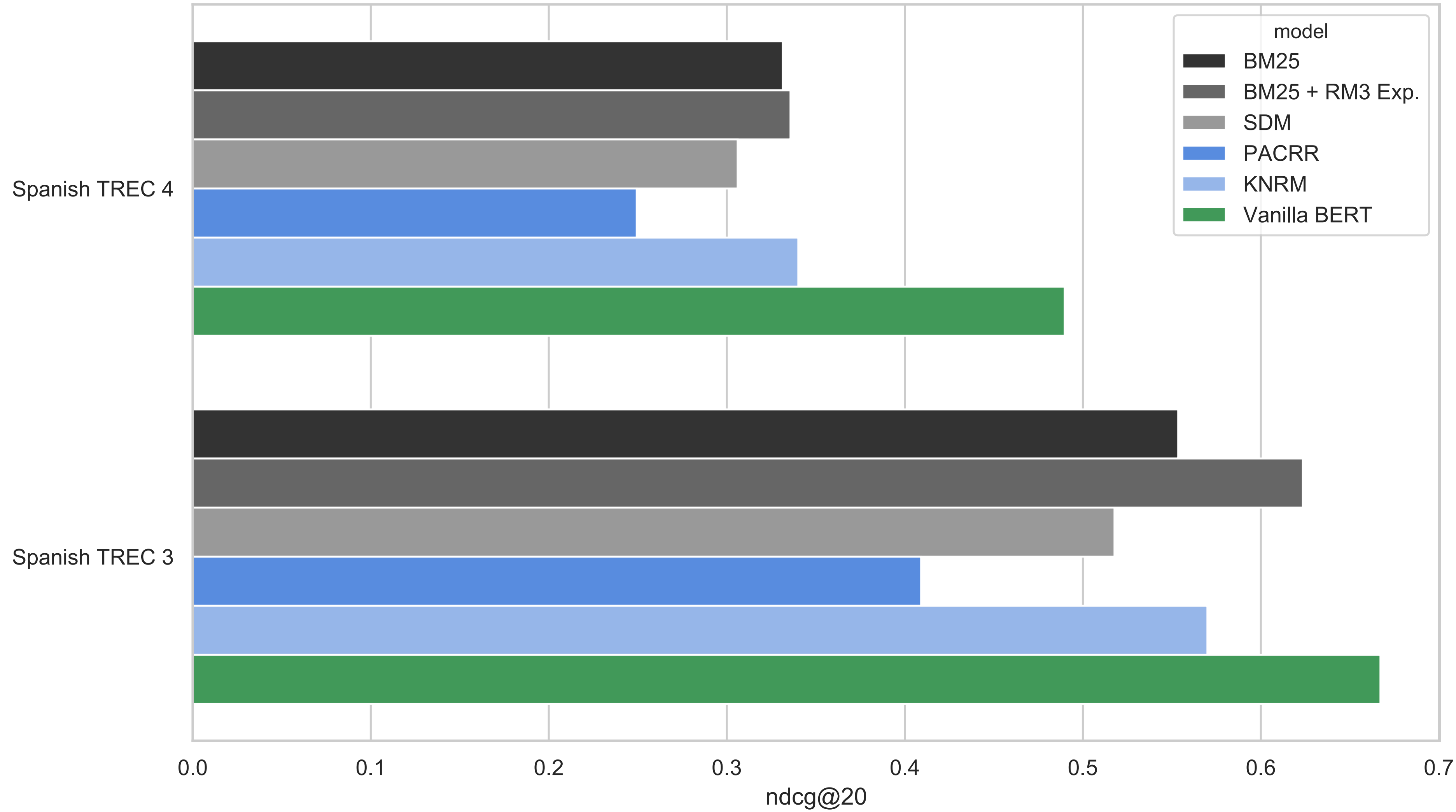
Traditional baseline methods: Probabilistic, Query Expansion, Query Likelihood



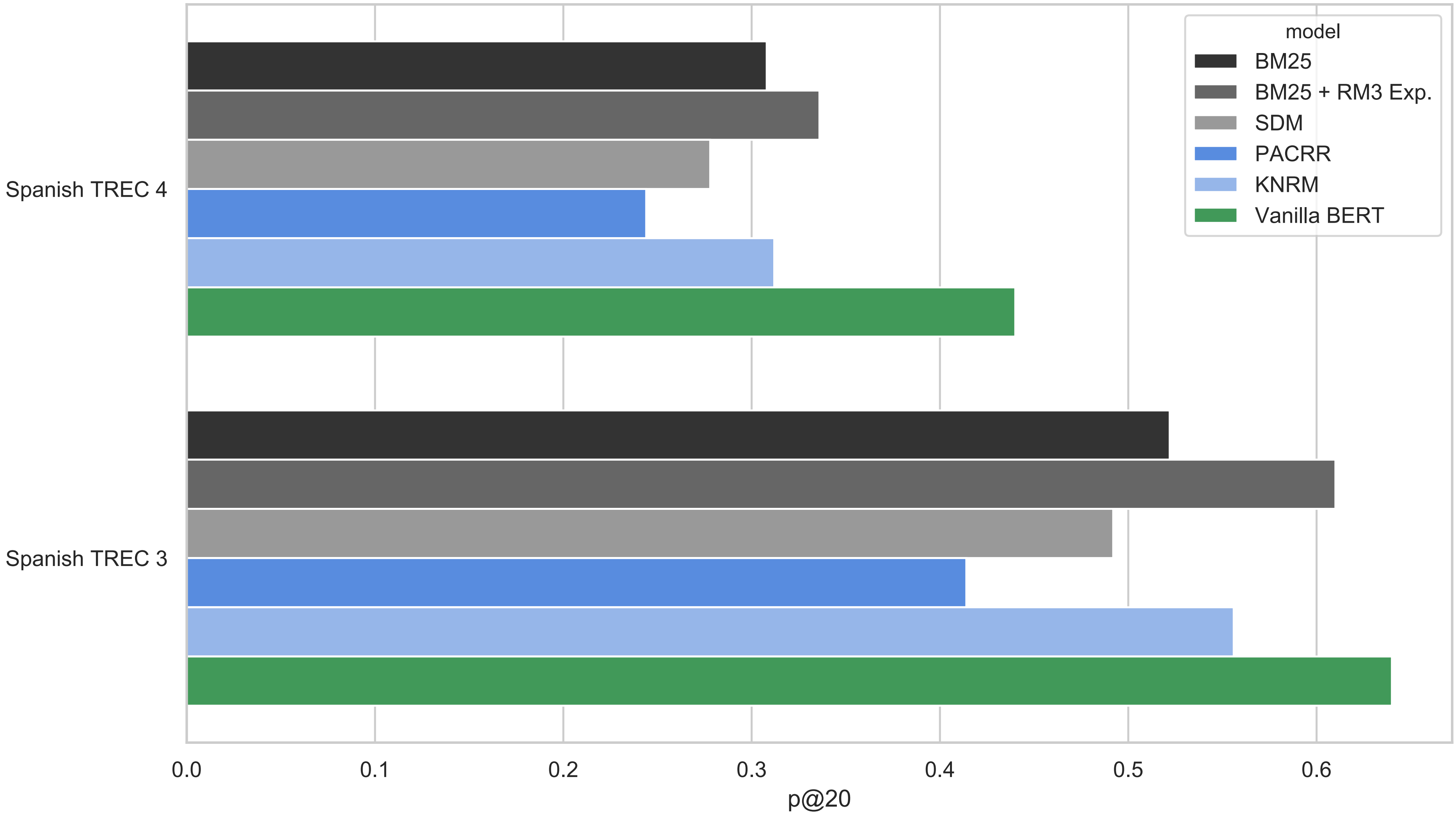
Baseline neural approaches:
PACRR (soft n-grams), KNRM (soft matching freq.)



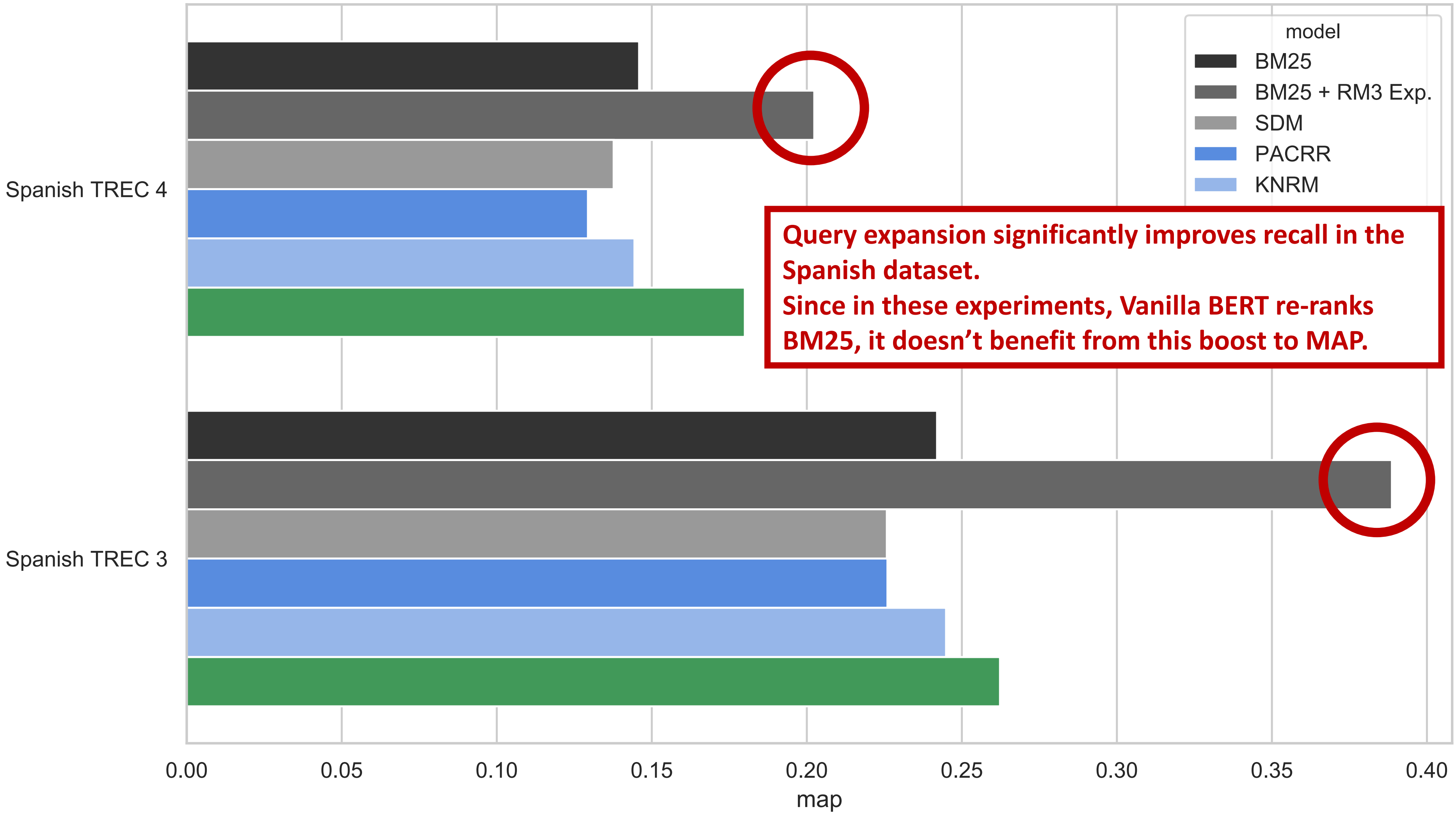
Multi-lingual Vanilla BERT significantly and consistently outperforms baseline rankers.



The trend holds across **metrics**… Precision @ 20



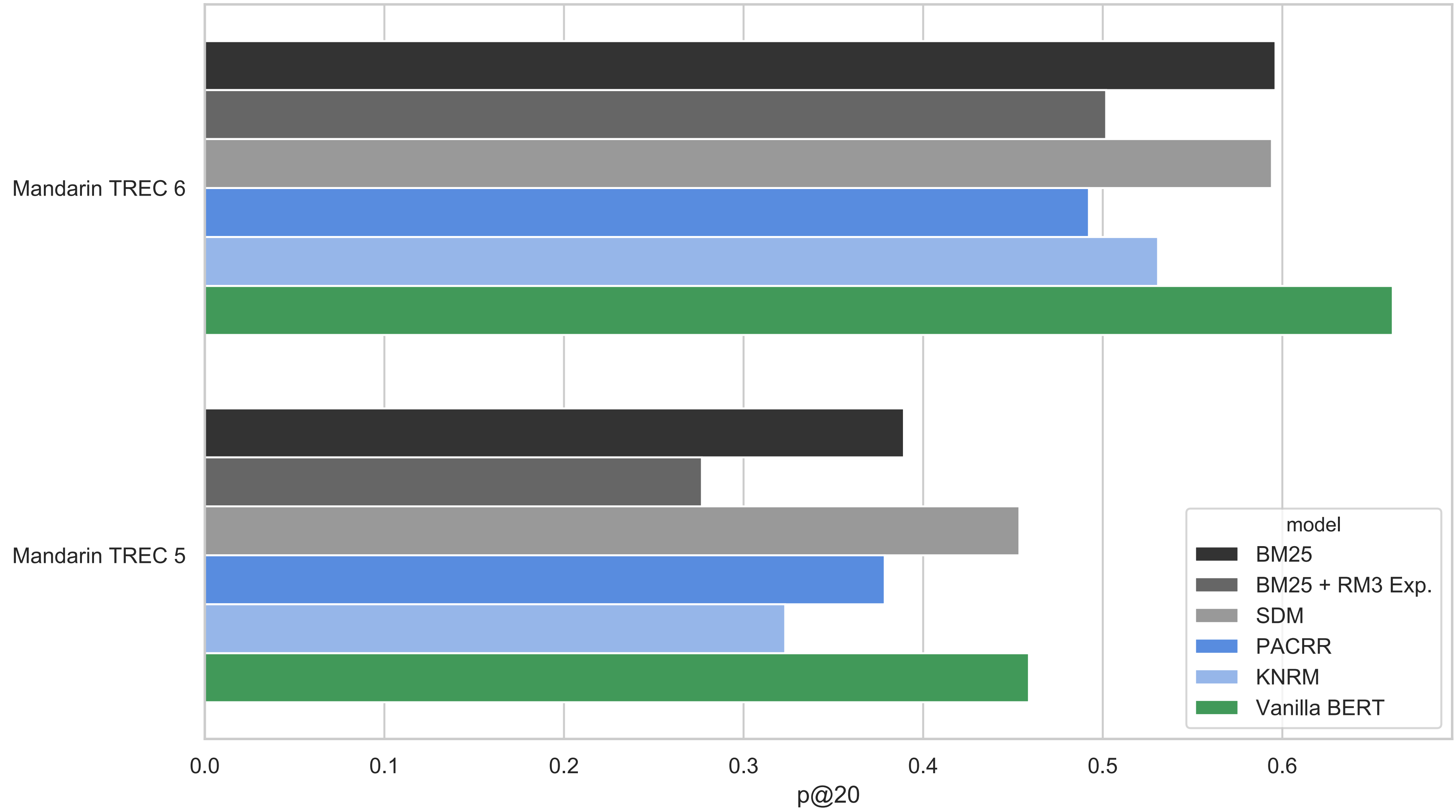
The trend holds across **metrics**… Mean Average Precision



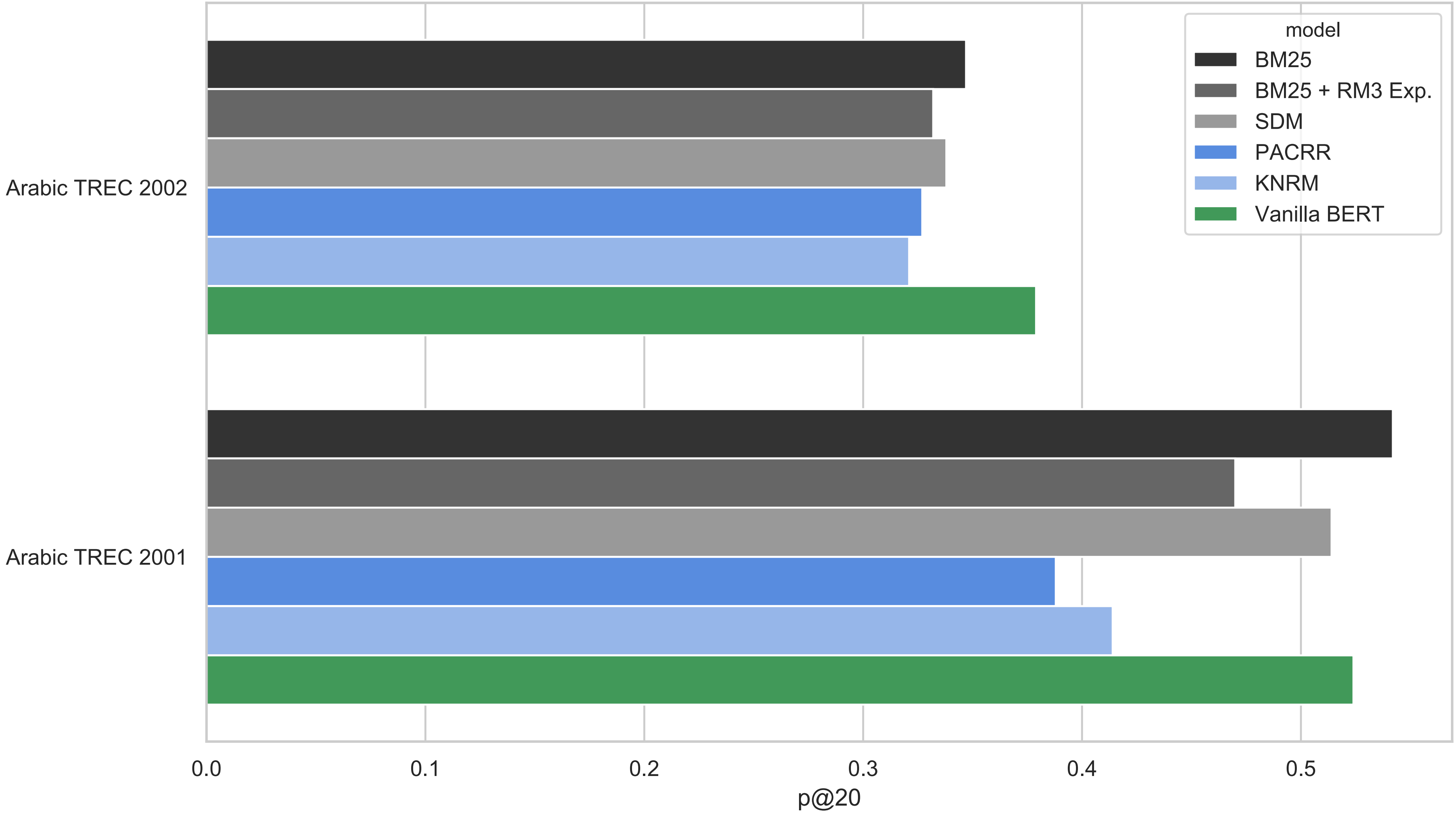
Query expansion significantly improves recall in the Spanish dataset.

Since in these experiments, Vanilla BERT re-ranks BM25, it doesn’t benefit from this boost to MAP.

The trend holds across **languages**… Mandarin

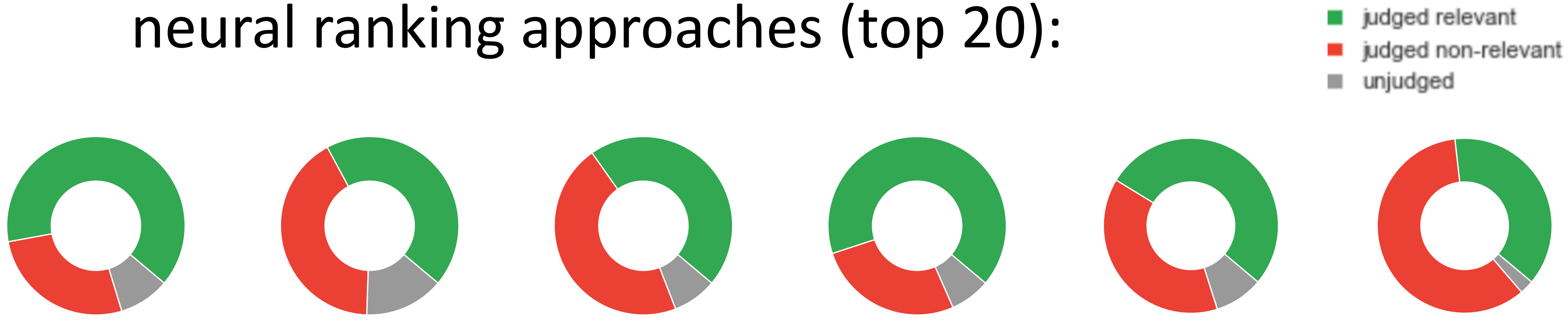


The trend holds across **languages**… Arabic



Observations:

- Traditional approaches vary wildly in their effectiveness across languages.
- Other neural approaches consistently under-perform (or perform comparably to) BM25.
- Multilingual Vanilla BERT significantly improves upon BM25 in almost all cases.
- The datasets are still effective for evaluating recent neural ranking approaches (top 20):

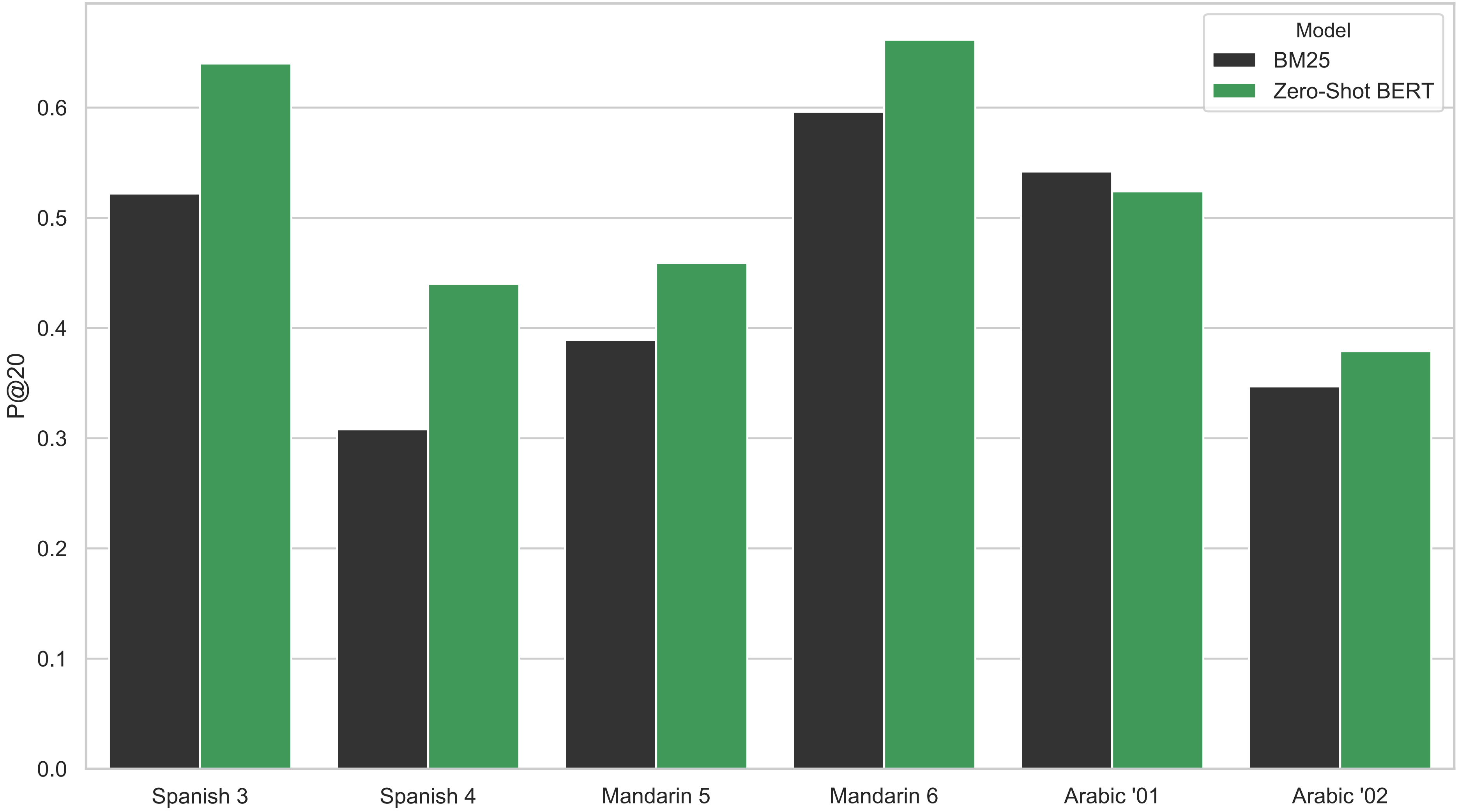


Spanish TREC 3 (90.8% judged)　Spanish TREC 4 (85.6% judged)　Mandarin TREC 5 (92.0% judged)　Mandarin TREC 6 (92.7% judged)　Arabic TREC 2001 (91.0% judged)　Arabic TREC 2002 (97.4% judged)

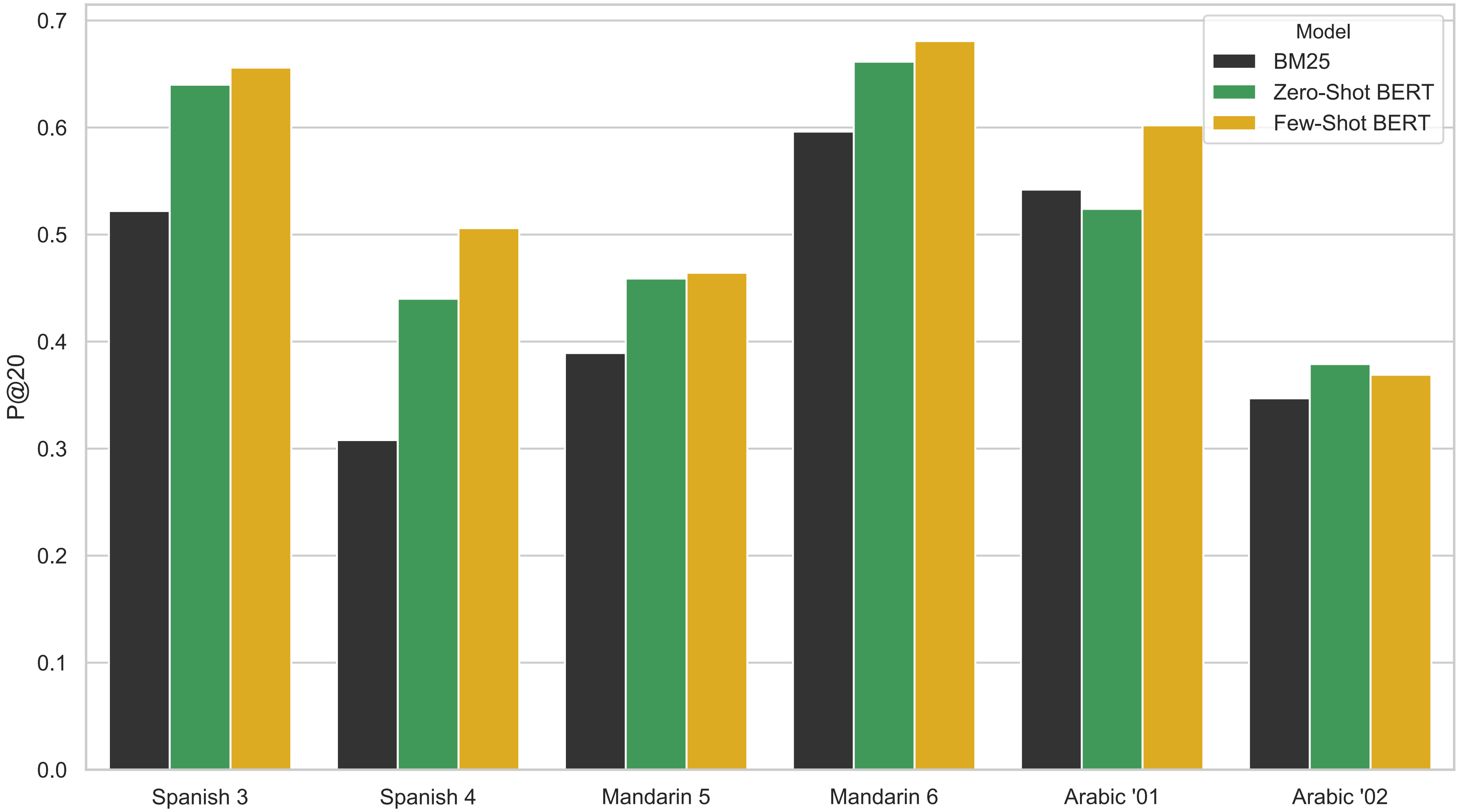
Does adding in-language data help?

- Experiment: **Few-Shot setting**
- Interleave examples from target language in training data (using other evaluation dataset for each language)

(Another view of the data from previous slides.)



Adding some in-language relevance samples often helps.



Future work

- We do not know if these models would perform better when trained on a suitable amount of in-language data.
- We do not know how well this works on low-resource languages.
- Our experiments used news documents.
  Are there considerations to make for other domains?

Concurrent work

- Peng Shi, Jimmy Lin. Cross-Lingual Relevance Transfer for Document Retrieval. arxiv:1911.02989
- Similar observations to ours

Cross-Lingual Relevance Transfer for Document Retrieval

Peng Shi and **Jimmy Lin**
David R. Cheriton School of Computer Science
University of Waterloo

Abstract

Recent work has shown the surprising ability of multi-lingual BERT to serve as a zero-shot cross-lingual transfer model for a number of language processing tasks. We combine this finding with a similarly-recently proposal on sentence-level relevance modeling for document retrieval to demonstrate the ability of multi-lingual BERT to transfer models of relevance across languages. Experiments on test collections in five different languages from diverse language families (Chinese, Arabic, French, Hindi, and Bengali) show that models trained with English data improve ranking quality, without any special processing, both for (non-English) mono-lingual retrieval as well as cross-lingual retrieval.

1 Introduction

Transformer models that have been pretrained on language modeling tasks such as BERT (Devlin et al., 2019) have led to many advances in diverse language processing tasks ranging from textual inference to sequence labeling. Interest in these models have also extended to search-related tasks such as retrieval-based question answering (Yang et al., 2019a), passage ranking (Nogueira and Cho, 2019), and document ranking (Yang et al., 2019b; MacAvaney et al., 2019; Yilmaz et al., 2019).

Our work builds on Yilmaz et al. (2019), who proposed a simple approach to document ranking that aggregates sentence-level evidence (based on BERT) with document-level evidence (based on traditional exact term-matching scores). Furthermore, they demonstrated that BERT models fine-tuned with passage-level relevance data can transfer across domains: surprisingly, fine-tuning on social media data is effective for relevance classification on newswire documents without any additional modifications.

Inspired by the work of Wu and Dredze (2019), who explored the cross-lingual potential of multilingual BERT (henceforth, mBERT for short) as a zero-shot language transfer model, we wondered if the techniques of Yilmaz et al. (2019) would transfer across languages in addition to transferring across domains. Supported by experiments in five different non-English languages from diverse language families (Chinese, Arabic, French, Hindi, and Bengali)—we find, perhaps unsurprisingly, the answer is *yes!*

The contribution of this work is empirical validation that the cross-domain relevance transfer work of Yilmaz et al. (2019) also works *cross-lingually* without any additional effort, for both mono-lingual retrieval in non-English languages as well as cross-lingual retrieval. We demonstrate robust increases in document retrieval effectiveness across diverse languages that come "for free".

2 Background and Approach

Our work adopts the standard formulation of document ranking: *given* a user query Q, the system's task is to produce a ranking of documents from a corpus that maximizes some ranking metric—in our case, average precision (AP). In the context of cross-lingual transfer learning, it is useful to precisely define the *source* language (the language of the training data) and the *target* language (the language in which inference is being applied). In our case, the source language is English. There are two variants of our retrieval task: In mono-lingual target language retrieval, both the query and the documents are in another language (for example, Bengali). In cross-lingual retrieval, the query and the documents are in different languages (for example, English queries, Bengali documents).

Following Wu and Dredze (2019), we use mBERT, which has been pretrained on concate-

Teaching a New Dog Old Tricks: Resurrecting Multilingual Retrieval Using Zero-shot Learning

Sean MacAvaney, Luca Soldaini, Nazli Goharian

sean@ir.cs.georgetown.edu · https://macavaney.us · @macavaney

- New neural ad-hoc ranking approaches can be trained cross-lingually
- Contextualized language models pre-trained on multiple natural languages can be fine-tuned on English relevance samples
- Multilingual collections still exhibit a high proportion of judged docs
- Additional relevance samples in target language can further help

BERT graphic from Jay Alammar, http://jalammar.github.io/illustrated-bert/

Extra Slides

BERT document encoding

Linear layer atop [CLS] to produce ranking score

```
[CLS] <all query toks> [SEP]
<batched doc toks> [SEP]
```

To work around maximum lengths imposed by BERT’s positional embeddings (512), we batch segments of the document (always with entire query).

Convenient Observation

Although contextualized language models require a considerable amount of text to train effectively,

you can just use the document collection for this purpose if another resource does not exist.

bert-base-multilingual-cased

- 12-layer, 768-hidden, 12-heads, 110M parameters
- ~120k WordPiece tokens (English BERT has ~30k)