Entity Retrieval from Multilingual Knowledge Graphs

Saher Esmeir
Bloomberg
London, United Kingdom
sesmeir2@bloomberg.net

Arthur Câmara[*]
Delft University of Technology
Delft, The Netherlands
A.BarbosaCamara@tudelft.nl

Edgar Meij
Bloomberg
London, United Kingdom
emeij@bloomberg.net

Abstract

Knowledge Graphs (KGs) are structured databases that capture real-world entities and their relationships. The task of entity retrieval from a KG aims at retrieving a ranked list of entities relevant to a given user query. While English-only entity retrieval has attracted considerable attention, user queries, as well as the information contained in the KG, may be represented in multiple—and possibly distinct—languages. Furthermore, KG content may vary between languages due to different information sources and points of view. Recent advances in language representation have enabled natural ways of bridging gaps between languages. In this paper, we, therefore, propose to utilise language models (LMs) and diverse entity representations to enable truly *multilingual entity retrieval*. We propose two approaches: (i) an array of monolingual retrievers and (ii) a single multilingual retriever trained using queries and documents in multiple languages. We show that while our approach is on par with the significantly more complex state-of-the-art method for the English task, it can be successfully applied to virtually any language with an LM. Furthermore, it allows languages to benefit from one another, yielding significantly better performance, both for low- and high-resource languages.

1 Introduction

Knowledge graphs (**KG**s) are key for many search applications. Consider, for example, the user query "chess world champions". Modern search engines often present users with a list of world chess champions along with additional facts encoded as relations in a KG. The queries themselves, as well as the information contained in a KG, may be represented in multiple—and possibly distinct—languages. This poses a challenge to traditional entity retrieval methods usually optimised for a single language. In this work, we aim to tackle the task of *multilingual entity retrieval*: given a query in any language, and a KG holding data in multiple languages, retrieve a ranked list of relevant entities.

The task of entity retrieval, when both the query and KG are in English, is well-studied. Recent years have seen remarkable progress, resulting in over 20% improvement on DBpedia-Entity v2 (**`DE-v2`**), the standard test collection for the task (Hasibi et al., 2017). Works like ESim (Gerritse et al., 2020) and KEWER (Nikolaev and Kotov, 2020) utilised word embedding techniques to represent entities and user queries in the same latent space. Meanwhile, EM-BERT (Gerritse et al., 2022) combines a powerful entity extractor that enhances user queries with a pre-trained language model (LM), fine-tuned on another ranking task, to establish a new state-of-the-art. These methods, however, operated on a single language at a time and were not studied in a multilingual setting.

While `DE-v2` is an English-only collection, Wikipedia and DBpedia (Auer et al., 2007) provide a unique opportunity: because the contributors to each language edition come from different backgrounds and have different views, we often see rich and diverse entity representations that go well beyond word-for-word translation. Moreover, many entities are available only in some chapters but not in others (see Appendix G for examples). Thanks to its graph-based nature, DBpedia facilitates mapping between languages and different entities representing the same subject. This, in turn, allows us to build rich, multilingual representations.

Expanding `DE-v2` to multiple languages, however, carries several risks. The collection was developed based on English DBpedia; therefore, its pooling stage uses keyword-based retrievers optimised for English. Moreover, annotators were only presented with English content. In this paper, we discuss these challenges through example queries

[*] Research conducted when the author was doing an internship at Bloomberg.

Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL), pages 1 - 15
December 8, 2022 ©2022 Association for Computational Linguistics

and stress the importance of building a truly multilingual collection end-to-end.

To address the task of multilingual entity retrieval, we introduce BERTE, a multi- and crosslingual entity ranking framework. Despite its simple design and its flexibility to use any LM out of the box, it is comparable to the state-of-the-art on DE−v2 in its original English form and thrives in a variety of languages, including Spanish, Arabic, and Hebrew. Furthermore, our experiments show that BERTE can benefit greatly from combining information from multiple languages to boost its performance, establishing a new state-of-the-art for a large subset of the queries.

The main contributions of our work are threefold: (i) A novel and simple yet effective entity retriever for the monolingual setup; (ii) A system for multilingual entity retrieval; and (iii) A systematic way to extend DE−v2 to multiple languages accompanied with a set of strong baseline results.

2 Background and Related Work

Entity Retrieval While earlier works on retrieving entities from a KG relied heavily on the graph's structure (Ciglan et al., 2012; Neumayer et al., 2012; Nikolaev et al., 2016), recent works have shown a tendency towards using graph *embeddings* instead (Gerritse et al., 2020; Nikolaev and Kotov, 2020; Komamizu, 2020; Jameel et al., 2017; Liu et al., 2019; Naseri et al., 2018). These methods generally implement a keyword-based first-stage ranker, such as BM25 (Robertson et al., 1995) and then a learned reranker. Meanwhile, the current state-of-the-art on DE−v2, EM-BERT, relies on a state-of-the-art entity extractor (van Hulst et al., 2020) to add textual representations of entities to user queries, combined with a pretrained LM, which already entails part of the domain knowledge (Petroni et al., 2019). To do so, they apply a linear transformation with aligned entity and word piece vectors, similar to E-BERT (Poerner et al., 2020). EM-BERT also uses a two-stage fine-tuning procedure: First, on MS MARCO (Campos et al., 2016), a large passage ranking dataset. Then, the model is further fine-tuned on the actual query-entity pairs from the training set of DE−v2. While powerful, this approach is restricted due to its requirements. On the other hand, our work achieves similar performance in English without relying on an entity extractor, pre-calculated entity embeddings, or additional large-scale fine-tuning. It is, therefore, much easier to extend to other languages

Entity Linking Entity linking aims at identifying and assigning entity mentions in a piece of text (FitzGerald et al., 2021; van Hulst et al., 2020; Shen et al., 2021). GENRE (De Cao et al., 2020), for instance, uses BART (Lewis et al., 2019) and Beam Search to generate names of entities. On the other hand, BLINK (Wu et al., 2020) uses a two-stage zero-shot linking algorithm, where a very short textual description represents each entity. While methods could be shared between both tasks, here we focus purely on a retrieval task, where user queries are formed by a specific information need.

Neural Information Retrieval Neural methods have been shown to improve significantly keyword-based retrievers in a wide range of tasks (Mitra and Craswell, 2018), including ad-hoc retrieval (Nogueira et al., 2019; Dai and Callan, 2020; Yu et al., 2021; Nogueira and Cho, 2019; Akkalyoncu Yilmaz et al., 2019; MacAvaney et al., 2019; Câmara and Hauff, 2020), question answering (Yu et al., 2021), semantic reasoning (Xu et al., 2020), and link prediction (Daza et al., 2021). Several *Retrievers* that ditch the initial keyword-based ranking in favour of an end-to-end approach have recently been proposed (Khattab and Zaharia, 2020; Karpukhin et al., 2020; Xiong et al., 2020; Formal et al., 2021). While we do not tackle this problem in this paper, we acknowledge that it is a natural direction for future work on entity retrieval.

Knowledge Graph Embeddings Graph embeddings have evolved greatly. With the introduction of Graph Neural Networks (Wu et al., 2021), methods like TransH (Wang et al., 2014), HINGE (Rosso et al., 2020) and StarE (Galkin et al., 2020) rose quickly in popularity. With the inclusion of LMs, even more powerful methods appeared (Poerner et al., 2020; Broscheit, 2019; Liu et al., 2020a). These methods are usually focused on general-purpose embeddings and then utilised by entity retrieval systems, such as EM-BERT.

Multilingual and Crosslingual Retrieval A system is considered *multilingual* when information can be retrieved in two or more languages. Meanwhile, *crosslingual* systems enable queries to benefit from information sources in different languages, even if not explicitly trained in these (Peters et al., 2012; Conneau and Lample, 2019). For example, Nair et al. (2020) use neural methods to translate queries in context, while Litschko et al. (2018) employ an unsupervised approach with multilingual embeddings. Recently, van der Heijden et al. (2021) studied how meta-learning can help with multilingual and crosslingual text classifications using a version of XLM. (Conneau and Lample, 2019). These multilingual models, even before the fine-tuning stage, Even before the fine-tuning stage, these multilingual models already have crosslingual capabilities thanks to the multilingual sources presented during pretraining (Muller et al., 2021). Winata et al. (2021) studied the applicability of few-shot learning in a multilingual setting on natural language understanding tasks. They demonstrated that given a few examples in English, the model could perform better than random in other unseen languages. Zhang et al. (2021) presented Mr. TYDI, a multilingual collection for mono-lingual retrieval in multiple languages, designed to evaluate ranking with learned representations and zero-shot results.

Multilingual Entity Retrieval The task of multilingual entity retrieval is somewhat unexplored, given the lack of a truly multilingual benchmark. De Cao et al. (2021) presented mGENRE for multilingual entity linking. It matches its input against generated entity names from multiple languages, which allows for exploiting language connections and the richness of Wikipedia. Similarly, Botha et al. (2020) provided a method for linking entities in 100 languages using BERT encoders. Tsai and Roth (2016) addressed the related task of crosslingual Wikification, where the goal is to find the English title given a foreign mention.

3 Multilingual Entity Retrieval

To tackle the entity retrieval task, we follow a standard two-staged approach: we first use a keyword-based method to retrieve a set of entities and then rerank them. Both steps rely exclusively on textual information extracted from the KG. Similar to the guidelines in DE-v2, each entity representation is composed by concatenating its direct literal attributes.[1] For an entity e, with n_a textual attributes, its representation e_d is defined as $[a_t, a_{la}, a_1, \ldots, a_{n_a}]$, where a_t is the title, a_{la} is the long abstract and a_i is the i^{th} attribute. Appendix A provides an overview of our proposed system.

For the first-stage retrieval, we use the well-established and language-agnostic BM25. It scores documents in relation to a query based on term frequency, document frequency, document length and term saturation. Where possible, and to allow a fair comparison with earlier works, we use officially available run files[2] of BM25 or BM25F$_{ca}$.[3]

3.1 Neural Reranker

BERT-based rankers are generally classified as cross or bi-encoders. The former concatenate queries and documents to form a single input to the base LM (Nogueira and Cho, 2019; MacAvaney et al., 2019), while the latter computes query and entity embeddings separately and uses the similarity between their embeddings to estimate relevance (Hofstätter et al., 2020; Karpukhin et al., 2020). Here we opt for bi-encoders, given their ability to compute document embeddings offline.

In practical terms, given a query q (up to $n_q = 32$ tokens) and an entity textual representation e (up to $n_e = 200$ tokens), we score the pair using the dot product of their embeddings $E_q \cdot E_e$, where:

$$E_q = W^T \cdot \text{BERT}(\text{“}[Q]q_0q_1, \ldots q_{n_q}\text{”}), \quad (1)$$

$$E_e = W^T \cdot \text{BERT}(\text{“}[D]e_0e_1, \ldots e_{n_e}\text{”}). \quad (2)$$

While the score from the dot product is sufficient to rerank, we follow the common practice in Entity Retrieval (Gerritse et al., 2020; Nikolaev and Kotov, 2020) of mixing the LM-based score with the normalised scores of the first-stage retriever:

$BERT\text{E}(q,e) = \beta \cdot BM25(q,e) + (1-\beta) \cdot (E_q \cdot E_e)$

3.2 Monolingual Entity Reranking

The wide adoption of LMs in English NLP led to the introduction of many language-specific models, such as ArBERT (Abdul-Mageed et al., 2021) for Arabic, AlephBERT (Seker et al., 2021) for Hebrew, and Berto (Cañete et al., 2020) for Spanish. Recall that our first-stage retriever uses the language agnostic BM25. In the monolingual setup, with queries and documents in the same language l, we first retrieve entities covered in the l subgraph of the KG and then rerank using a BERT model pretrained on l and fine-tuned on triples $\langle q, e^+, e^- \rangle$ built from that subgraph. We refer to this version as BERTE$_l$. While including the structural components of the KG can be useful, we hypothesise that fine-tuning BERT using queries and textual data of entities is sufficient. Beyond that, it has been shown that a pretrained BERT model already has implicit domain knowledge (Bouraoui et al., 2020; Wang et al., 2020; Petroni et al., 2019).

[1] Unlike DE-v2, we use flat, unfielded documents.

[2] The run file provides, for each query, a scored list of 1000 entities retrieved by a keyword-based model.

[3] A fine-tuned version that uses fielded documents.

3.3 Entity Retrieval by Query Translation

Given a multilingual KG and a query in a non-English language l, a system could **M**achine **T**ranslation (MT) to obtain an English version of the query and then feed it to the English BERTE$_{en}$. It then utilises the graph to map the ranked entities back to l, if they exist.[4] We refer to this query translation method in our experiments as qtBERTE$_{en}$.

Due to its simplicity, qtBERTE$_{en}$ suffers from several shortcomings when used on a multilingual KG, such as DBpedia. Mainly, it is restricted to content in English, even if the graph holds information in multiple languages, and entities without English representation or entities with additional essential information in other languages will be missed. *This forces the English point of view on all users and ignores other, potentially more diverse, viewpoints.* Another issue with qtBERTE$_{en}$ is its reliance on MT. Despite the impressive progress, MT still needs improvement, especially for low-resource languages, with named entities presenting a significant challenge (Li et al., 2021). Moreover, in gender-marking languages, like Arabic, Hebrew and Spanish, gender hints will be lost.

3.4 Multilingual Entity Retrieval System

BERTE$_l$, by design, supports a single language. To handle queries and entities in multiple languages, an array of BERTE$_l$ models is needed, each of which uses a different LM. However, training an LM for a new language requires large amounts of data and significant computing power, limiting advances in NLP to a small subset of languages (Joshi et al., 2020). Moreover, fine-tuning and storing a model for each language is prohibitively expensive when the task involves more than a handful of languages. To overcome these challenges, multilingual LMs such as mBERT (Devlin et al., 2019) and mLUKE (Ri et al., 2021) were proposed, with the idea of training a single model for many languages.

multiBERTE, our proposed multilingual ranker, can handle any language supported by its base multilingual LM. We explore two approaches for multiBERTE: multi$_{\text{en}}$BERTE, which fine-tunes the multilingual BERT model using English data only, and a *few-shot* approach, multi$_{\text{few}}$BERTE, where training data from a few languages is concatenated. In the latter, the model has no explicit knowledge of what language it will use and only has a few training samples in each (Longpre et al., 2021).

Given training data in a language l, we can extend it to another language $l^{\times}$ by: (i) machine-translating the queries; and (ii) using the entity documents generated from the subgraph of $l^{\times}$.

Figure 1 compares the workflows of BERTE$_l$ and multiBERTE. The former only sees data in one language, both when pretrained and fine-tuned. Therefore, an array of BERTE$_l$ models is needed in a multilingual setup. multiBERTE, on the other hand, is pretrained with over 100 languages and can handle pairs in any of these languages, even if fine-tuned only on a subset of them.

3.5 Mixture of Language Rankers

Given a query written in a language l, qtBERTE$_{en}$ searches the English subgraph only, and its results are limited to entities that can be mapped to l. BERTE$_l$, on the other hand, considers only entities represented in l and uses the textual representation available in l in both stages. Consequently, information in other languages is not utilised. multi$_{\text{few}}$BERTE can take advantage of content from multiple sources during the fine-tuning stage but uses only l for retrieval.

We believe that, by mixing multiple models during retrieval, we can further benefit from the unique traits of individual subgraphs while diminishing biases that may have been encoded due to reliance on a single source. One option is to concatenate the different textual representations into a single multilingual document and use the combined document for fine-tuning and scoring. This approach will only work with multiBERTE. Even then, the limited document size most LMs can handle presents a barrier. An alternative is to translate the query to multiple languages and retrieve a scored list of entities for each language. We denote this approach of using multiple retrievers by adding the superscript L_{mix} to the model name. Formally, let l be the target language and L_{mix} be the set of additional languages we want to blend in, the mixed score is:

$$\text{BERTE}_l^{L_{\text{mix}}}(q,e) = \sum_{l^{\oplus} \in \{l\} \cup L_{\text{mix}}} \mu_{l^{\oplus}} \cdot \text{BERTE}_{l^{\oplus}}(q,e). \quad (3)$$

Note that BERTE$_{l^{\oplus}}$ could be a different BERTE$_l$ model for each language or a single multiBERTE model shared between all languages. $\mu_{l^{\oplus}}$, the weight each language gets in the final score can be learned based on factors including geographical location, language similarity, or user preference. In this work, we assign a fixed weight of $\mu_l = 0.75$ to

[4] Section 4 shows how the DBpedia entity mapping works.

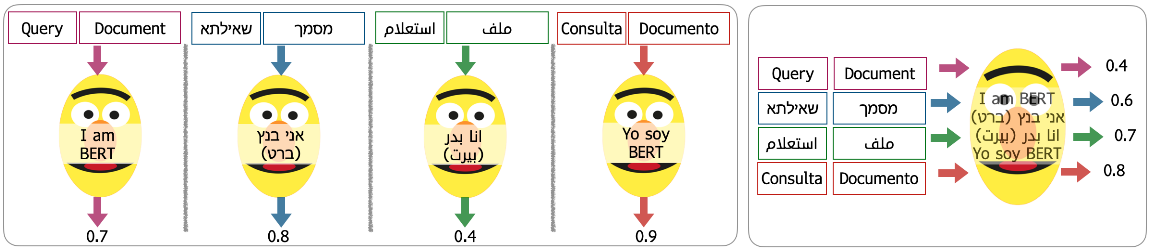

Figure 1: We consider two architectures for a multilingual retrieval system: BERTE$_l$, a collection of monolingual retrievers (left) and a single multilingual model, multiBERTE, trained using query-document pairs from multiple languages (right).

the target language ranker and split the remaining weight equally between the rest. For example, if the target language is Arabic, BERTE$_{ar}^{\{en\}}$ will be a mixture of BERTE$_{ar}$ and BERTE$_{en}$. The Arabic and English versions of the queries are used. The weight of the BERTE$_{ar}$ score will be $\mu = 0.75$ and the weight of BERTE$_{en}$ will be 0.25.

Appendix B compares the various configurations in our multilingual retrieval system.

4 Empirical Evaluation

We conducted a series of experiments on DE−v2 to analyse our proposed approaches. We also used DE−v2's 5-fold train-test split to allow comparison with previous works. β, the weight given to the first-stage retriever, is fine-tuned using a validation set (one training fold). We found $\beta = 0.75$ to work best for English and used it across all experiments.

In each language l, we adopt the same procedure when training the respective BERTE$_l$ model. For every training query q and relevant entity e^+, we generate 10 triplets of the form $\langle q, e^+, e^- \rangle$, where e^- is a randomly drawn judged non-relevant entity for q. We use a pairwise softmax cross-entropy loss, AdamW optimiser, with a learning rate of $1e^{-6}$, and train for 20,000 steps, with a batch size of 32. The embedding vectors are of size 128.

We first evaluate BERTE on the original English collection and the Arabic subset of DE−v2, the only publicly available non-English resource for the task. We then discuss how to extend DE−v2 to other languages systematically and evaluate BERTE$_l$ (monolingual LMs) and multiBERTE (multilingual LMs) on the complete set of queries, machine-translated to Spanish, Arabic, and Hebrew. Finally, we demonstrate how English can benefit from other languages. Note that we optimise for English only and fix $\beta = 0.75$ for all experiments and languages. Optimising β per language will likely further improve results.

Table 1: Reranking results. Statistically significant improvements (paired t-test with $\alpha = 0.05$) over ESim and KEWER are indicated by ($\star$) and (†) respectively.

Model	nDCG$_{10}$	nDCG$_{100}$	MAP
BM25F$_{ca}$	0.461	0.551	0.380
KEWER	0.483	0.560	0.396
ESim	0.487	0.572	0.403
EM-BERT	$0.541^{\star\dagger}$	$0.604^{\star\dagger}$	-
BERTE$_{en}$	$0.525^{\star\dagger}$	$0.602^{\star\dagger}$	$0.433^{\star\dagger}$

4.1 Evaluating BERTE on English

DE−v2 comes with a set of baseline results. The official metrics are nDCG (Normalized Discounted Cumulative Gain) at 10 and 100. Similar to other works, we also report MAP (Mean Average Precision) at 1,000. We utilise the recently introduced embedding-based techniques KEWER and ESim, as well as EM-BERT, which uses LMs, as baselines.[5] We reproduced the baselines reported results using their published runs, if available. Table 1 shows the overall results. [6] Our proposed BERTE$_{en}$ and the current state-of-the-art EM-BERT significantly outperform the other methods (paired t-test with $\alpha = 0.05$). Between them, the differences in nDCG are statistically insignificant. We believe, however, that BERTE$_{en}$ is preferred, even in a monolingual setting, for the following reasons: (i) it uses a smaller LM (BERT-base vs BERT-large); (ii) it does not require additional annotated data and instead has a single fine-tuning step; (iii) it does not depend on the availability of entity embeddings and entity extractors; (iv) it re-ranks directly from BM25 instead of ESim.

Our main focus in this work, nevertheless, is the multilingual setting. We, therefore, use Appendix

[5] We use the model with the best reported overall result for each: BM25F$_{ca}$+KEWER, BM25F$_{ca}$+ESim$_{CG}$, and EM-BERT with GEEER and dual fine-tuning.

[6] Note that KEWER used a custom 5-fold split for cross-validation. MAP at 1000 could not be reported for EM-BERT because the run files are limited to 100 results.

C to dive deeper into the differences between the different methods in the English setup and show that an even better result can be achieved by combining them. In Appendix D, we also present insights from sample query analysis.

4.2 Evaluating BERTE on Arabic

Esmeir (2021) has recently used human-translators to extend DE-v2 to Arabic. Only 139 queries with sufficient relevant entities in Arabic were included. Along the translations, two baseline results were reported: BM25 and SERAG, an adaption of KEWER to Arabic.

Table 2: Reranking results on the Arabic collection. Significant improvements over SERAG and BM25 are indicated by ($\star$) and (†) respectively.

Model	nDCG$_{10}$	nDCG$_{100}$	MAP
SERAG	0.226	0.303	0.183
BM25	0.273*	0.3482*	0.223*
BERTE$_{ar}$	**0.308***†	**0.382***†	**0.247***†

As shown in Table 2, our BM25 first stage is already enough to outperform SERAG. We believe that this is due to better document generation. BERTE$_{ar}$ model provides a further statistically significant improvement.

4.3 Extending DE-v2

DBpedia can be viewed as a large-scale multilingual KG (Lehmann et al., 2015). Each chapter holds structured content extracted from the corresponding Wikipedia edition. Inter-language entity-mapping files allow us to link entity URIs from one language to another. Given an entity in the graph, we can extract its multilingual counterparts using the owl#sameAs property. Figure 2 illustrates how this linking works for the entity representing "Ibn Khaldun". Appendix E provides coverage statistics of the DBpedia 2015 chapters we use.

While there is at most one Wikipedia article per topic per language, the content of the articles may vary across languages. Moreover, editors and administrators from different editions may have different points of view. They may also have access to different sources, only available in that specific language. Finally, different languages may encode different biases into the LM (Bartl et al., 2020). Consider, for example, the topic *"Mujaddara"*, a popular dish in several parts of the world. Examining the info-boxes in different languages, we found over ten different answers to where they originated (as of early 2022). A good retriever, therefore, will attempt to benefit from the *richness of DBpedia by considering information from multiple languages.*

In DE-v2, retrieving an entity that does not exist in English may hurt the results. First, only entities with English content were judged by the annotators. Other entities, even if relevant to the query, will not have a judgement and will default to non-relevancy. Second, placing a relevant but unjudged entity in a high-ranking position may push other judged relevant entities outside the top k and hurt the measured performance. We restrict the first-stage retrieval to entities available in English DBpedia to solve the latter. This step, however, should not be applied in the general case, where judgements are truly multilingual.

To translate the queries, we opted for MT. While human translation provides major benefits, MT allows it to scale to over 100 languages.[7]

4.4 Monolingual Language Models

We next study Spanish, Arabic, and Hebrew versions of DE-v2, using the set of machine-translated queries. Table 3 shows the results. We first consider models that use a single-language subgraph (top three lines), where the language-agnostic BM25 is provided as a baseline. Similar to the English results, a BERTE$_l$ model fine-tuned in the same language significantly outperforms BM25. Interestingly, qtBERTE$_{en}$, which uses English queries to search the English chapter of DE-v2 and maps the results back to the target language, outperforms BERTE$_l$.

While the "English-first" approach seems to perform better than searching specifically on a given language, we must treat this result with caution due to the "English nature" of DE-v2: (i) English is the largest chapter with the most coverage, (ii) entities were pooled using methods optimised for English, and (iii) the annotators had English Wikipedia in mind when judging the entities. In addition, recall that we have the original queries in English so BERTE$_{en}$ operated on optimal translations. We view this result as a strong baseline but not one that can generalise for the wealth of retrieval tasks in a truly multilingual universe.

Next, we investigate what happens if we *mix* the scores of BERTE$_l$ on a language l with those of

[7] We used http://translate.google.com and asked native speakers to verify that the output was generally in line with the input. We did not have to make any changes.

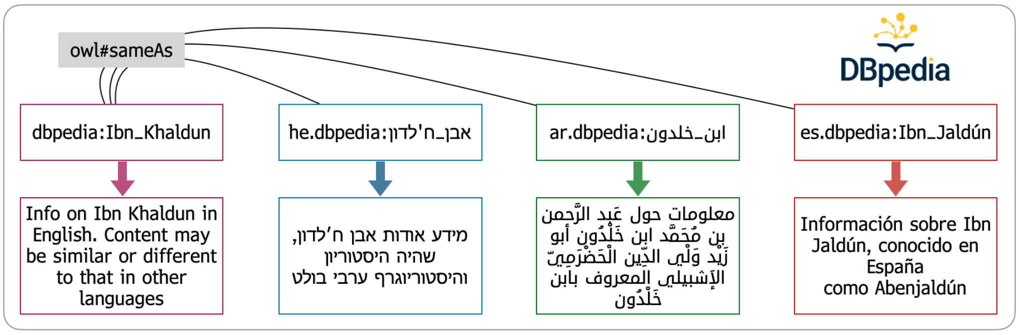
owl#sameAs
DBpedia
dbpedia:Ibn_Khaldun
he.dbpedia:אבן_ח'לדון
ar.dbpedia:ابن_خلدون
es.dbpedia:Ibn_Jaldún
Info on Ibn Khaldun in English. Content may be similar or different to that in other languages
מידע אודות אבן ח׳לדון, שהיה היסטוריון והיסטוריוגרף ערבי בולט
معلومات حول عَبد الرَّحمن بن مُحَمَّد ابن خَلْدُون أبو زَيْد وَلْي الدِّين الْحَضْرَمِيّ الإشبيلي المعروف بابن خَلْدُون
Información sobre Ibn Jaldún, conocido en España como Abenjaldún

Figure 2: Entity mapping between entities in DBpedia chapters. In this example, the English entity for “Ibn Khaldun” is mapped to the respective entities in Arabic, Hebrew and Spanish. The graph and content in each language may differ (the texts in the example are for illustration only).

Table 3: Reranking results in the multilingual setup. BERTE_l is trained solely on l. qtBERTE_{en} uses query translation from l to English, searches the English KG and maps the results to entities in the l graph. multi_{en}BERTE and $\text{multi}_{\text{few}}$BERTE are multilingual models, fine-tuned in English or few languages, respectively. $\text{multi}_{\text{en}}\text{BERTE}^{\{en\}}$ and $\text{multi}_{\text{few}}\text{BERTE}^{\{en\}}$ mix in the scores from qtBERTE_{en}. Best result in each column in **bold**. ↑ denotes significant improvements over the preceding line.

	English		Spanish		Arabic		Hebrew	
Model	**nDCG$_{10}$**	**nDCG$_{100}$**	**nDCG$_{10}$**	**nDCG$_{100}$**	**nDCG$_{10}$**	**nDCG$_{100}$**	**nDCG$_{10}$**	**nDCG$_{100}$**
BM25 (BM25F$_{ca}$ for English)	0.461	0.551	0.271	0.320	0.216	0.265	0.216	0.266
BERTE_l	0.525 $^{\uparrow}$	0.602 $^{\uparrow}$	0.299 $^{\uparrow}$	0.353 $^{\uparrow}$	0.242 $^{\uparrow}$	0.293 $^{\uparrow}$	0.238 $^{\uparrow}$	0.290 $^{\uparrow}$
qtBERTE_{en}	-	-	0.345 $^{\uparrow}$	0.446 $^{\uparrow}$	0.271 $^{\uparrow}$	0.349 $^{\uparrow}$	0.263 $^{\uparrow}$	0.345 $^{\uparrow}$
$\text{BERTE}_l{}^{\{en\}}$	-	-	0.472 $^{\uparrow}$	0.497 $^{\uparrow}$	0.421 $^{\uparrow}$	0.452 $^{\uparrow}$	0.415 $^{\uparrow}$	0.439 $^{\uparrow}$
multi_{en}BERTE	**0.530**	**0.608**	0.311	0.363	0.236	0.287	0.244	0.293
$\text{multi}_{\text{en}}\text{BERTE}^{\{en\}}$	-	-	**0.473** $^{\uparrow}$	**0.498** $^{\uparrow}$	0.420 $^{\uparrow}$	0.452 $^{\uparrow}$	0.414 $^{\uparrow}$	0.437 $^{\uparrow}$
$\text{multi}_{\text{few}}$BERTE	0.529	0.607	0.317	0.371	0.238	0.289	0.249	0.299
$\text{multi}_{\text{few}}\text{BERTE}^{\{en\}}$	-	-	**0.473** $^{\uparrow}$	0.497 $^{\uparrow}$	**0.422** $^{\uparrow}$	**0.453** $^{\uparrow}$	**0.417** $^{\uparrow}$	**0.440** $^{\uparrow}$

qtBERTE_{en}. The results in the fourth row of Table 3 shows that *mixing languages is highly beneficial.*

To illustrate that, consider the Hebrew version of the query “Chefs with a show on the Food Network”. “Julia Child” is the highest-scored entity from DBpedia Hebrew. While BM25 includes it in the first stage, BERTE_l ranks it outside the top 100. “the Food Network” was translated literally into *RESHET HAMAZON*, failing to identify the named entity. By mixing in the English score, “Julia Child” breaks into the top 10. In some cases, however, mixing scores is not strictly beneficial: The Spanish BERTE_l, for example, does better on the query “Madrid” without mixing English, perhaps unsurprisingly.

In our experiments, while models could leverage information from multiple languages, they return only entities covered in the language of the query. In some scenarios, however, it is useful to see entities that exist only in other languages. In Appendix G, we provide further insights into this setup and explain how BERTE can be adapted to handle it using score mixing.

4.5 Multilingual LMs

The results for the multiBERTE variants are presented in the last four rows of Table 3. Both multi_{en}BERTE (a multilingual model fine-tuned on English queries) and $\text{multi}_{\text{few}}$BERTE (a multilingual model fine-tuned with queries in multiple languages) exhibit comparable performance to the respective monolingual models, especially when mixing in the scores from English (rows 6 and 8), indicating that score blending offers an orthogonal advantage. While not reflected in the numbers, we believe that $\text{multi}_{\text{few}}$BERTE is the better choice, given its ability to incorporate knowledge from multiple languages and the fact that it can adapt quickly to a new language. We expect it to shine when using information from multiple languages is essential. We hope to collaborate with the community to build such truly multilingual collections.

Another advantage of multiBERTE variants is that they can be used for over 100 languages without having to fine-tune the models on these, thanks to the cross lingual capabilities of multilingual LMs,

Table 4: Results on queries with good coverage across languages. Sig. improvements are denoted by ($\star$).

Model	nDCG$_{10}$	nDCG$_{100}$	MAP
BERTE$_{en}$	0.515	0.616	0.434
BERTE$_{en}^{\{es,ar,he\}}$	**0.540***	**0.634***	**0.452***

leading to a strong baseline for many languages on the task. In Appendix F we provide results for six additional languages obtained following the same methodology as our main result. multi$_{\text{few}}$BERTE$^{\{\text{en}\}}$ was consistently the best performer, and its advantage over BM25 and multi$_{\text{few}}$BERTE was statistically significant.

While the multiBERTE variants offer apparent advantages, there are cases where BERTE$_l$ is necessary: (i) there are hundreds of languages that are not supported by existing multilingual LMs but have their own monolingual LM, and (ii) domain specific LMs, such as FinBERT (Liu et al., 2020b), were shown to be superior in many tasks.

4.6 English Benefits from Collaboration

Above, we demonstrated that mixing in the scores from BERTE$_{en}$ helps other languages. We next ask if English, the largest and richest chapter, can also benefit from the diverse coverage in other languages. To answer this, we mix Spanish, Arabic, and Hebrew scores to rerank English entities. We refer to this model as BERTE$_{en}^{\{es,ar,he\}}$.

When tested on the entire dataset, this approach did not yield any improvement. Error analysis, however, indicated that for queries where the Spanish, Arabic, and Hebrew runs of BM25 obtained a sufficient number of relevant entities, and the performance of BERTE$_{en}^{\{es,ar,he\}}$ on English improved. When entities do not exist in another language, or when their representation does not match the query textually, BM25 fails to retrieve them, negatively impacting the corresponding BERTE$_l$ and subsequently BERTE$_{en}^{\{es,ar,he\}}$. We, therefore, focus on a subset of the queries with good performance of BM25 in the other three languages. More formally, we calculate the optimal nDCG$_{100}$ of the first-stage retrieval, which provides an upper bound on reranking performance. Queries with a score of 0.3 or more in all languages are kept, resulting in a subset of 113 queries. Table 4 lists the results for this subset showing that the performance of BERTE$_{en}^{\{es,ar,he\}}$ is significantly better than BERTE$_{en}$. Consider, for example, the query "Chess world champions". The query is about a global topic with coverage in many languages. BERTE$_{en}$ listed 4 relevant names in its top 10 results. With the help of other languages, this number increased to 6. This demonstrates that *even high resource languages can benefit from multilingual retrieval*. Instead of pre-selecting a subset of queries, in the future, we plan to apply a meta-learner to decide which languages a query should use automatically.

5 Conclusion

In this paper, we introduced BERTE, a highly effective multilingual entity retrieval system. We showed that in a monolingual environment, it is on par with current state-of-the-art methods on DE−v2 despite being simpler and requiring less data. We then explored the multilingual setup, where both the graph and the queries may be presented in multiple languages. We proposed a systematic way to extend DE−v2 beyond English and discussed the risks of such approach. We believe it is vital for the community to curate truly multilingual collections that come from different sources and involve native speakers. To address the multilingual retrieval task, we considered both a collection of monolingual models and a single multilingual one. We showed that combining the scores from different languages significantly boosts the performance of low and high-resourced languages.

Our work can enable many downstream tasks. Consider, for example, a virtual assistant answering questions in Arabic about a topic covered mainly in the English edition of Wikipedia or an English speaker analyst covering a multi-national company interested in taking diverse points of view coming from content in different languages. In both cases, BERTE allows handling queries and KGs in multiple languages.

We hope that this work opens interesting avenues of research. As discussed in Appendix C, the improvements brought by BERTE are orthogonal to those by the other state-of-the-art method, EMBERT. Therefore, we hope that combining each method's contributions will establish a new state-of-the-art for the English task. On the multilingual front, works that adapt the retrieval task to user preference, such as language, region, or past actions, may benefit from the flexibility of BERTE in combining different sources.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Applying BERT to document retrieval with birch. 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): System Demonstrations*, pages 19–24, Hong Kong, China. Association for Computational Linguistics.

S. Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. Dbpedia: A nucleus for a web of open data. In *ISWC/ASWC*.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Jan A. Botha, Zifei Shan, and Daniel Gillick. 2020. Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845, Online. Association for Computational Linguistics.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7456–7463.

Samuel Broscheit. 2019. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 677–685, Hong Kong, China. Association for Computational Linguistics.

Arthur Câmara and Claudia Hauff. 2020. Diagnosing bert with retrieval heuristics. In *Advances in Information Retrieval*, pages 605–618, Cham. Springer International Publishing.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Marek Ciglan, Kjetil Nørvåg, and Ladislav Hluchý. 2012. The SemSets model for ad-hoc semantic list search. In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, pages 131–140, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*.

Zhuyun Dai and Jamie Callan. 2020. Context-aware document term weighting for ad-hoc search. In *Proceedings of The Web Conference 2020*, WWW '20, page 1897–1907, New York, NY, USA. Association for Computing Machinery.

Daniel Daza, Michael Cochez, and Paul Groth. 2021. Inductive entity representations from text via link prediction. In *Proceedings of the Web Conference 2021*, WWW '21, page 798–808, New York, NY, USA. Association for Computing Machinery.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2021. Multilingual autoregressive entity linking. *arXiv preprint arXiv:2103.12528*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Saher Esmeir. 2021. SERAG: Semantic entity retrieval from Arabic knowledge graphs. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 219–225, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nicholas FitzGerald, Dan Bikel, Jan Botha, Daniel Gillick, Tom Kwiatkowski, and Andrew McCallum. 2021. MOLEMAN: Mention-only linking of entities with a mention annotation network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 278–285, Online. Association for Computational Linguistics.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Mikhail Galkin, Priyansh Trivedi, Gaurav Maheshwari, Ricardo Usbeck, and Jens Lehmann. 2020. Message passing for Hyper-Relational knowledge graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7346–7359. Association for Computational Linguistics.

Emma Gerritse, Faegheh Hasibi, and Arjen De Vries. 2022. Entity-aware Transformers for Entity Search. In *Proc. of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22.

Emma J. Gerritse, Faegheh Hasibi, and Arjen P. de Vries. 2020. Graph-embedding empowered entity retrieval. *Advances in Information Retrieval*, pages 97–110.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1265–1268. ACM.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*.

Shoaib Jameel, Zied Bouraoui, and Steven Schockaert. 2017. MEmbER: Max-Margin based embeddings for entity retrieval. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 783–792, New York, NY, USA. Association for Computing Machinery.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Omar Khattab and Matei Zaharia. 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.

Takahiro Komamizu. 2020. Random walk-based entity representation learning and re-ranking for entity search. *Knowledge and Information Systems*, 62:2989–3013.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Panpan Li, Mengxiang Wang, and Jian Wang. 2021. Named entity translation method based on machine translation lexicon. *Neural Computing and Applications*, 33:1–9.

Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. 2018. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1253–1256.

Liyuan Liu, Jingbo Shang, and Jiawei Han. 2019. Arabic named entity recognition: What works and what's next. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 60–67, Florence, Italy. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020a. K-BERT: Enabling language representation with knowledge graph. *AAAI*, 34(03):2901–2908.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020b. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization. Special Track on AI in FinTech.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. 2019. CEDR: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR'19, pages 1101–1104, New York, NY, USA. Association for Computing Machinery.

Bhaskar Mitra and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Suraj Nair, Petra Galuscakova, and Douglas W Oard. 2020. Combining contextualized and non-contextualized query translations to improve clir. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1581–1584.

Shahrzad Naseri, John Foley, J. Allan, and Brendan T. O'Connor. 2018. Exploring summary-expanded entity embeddings for entity retrieval. In *CIKM Workshops*.

Robert Neumayer, Krisztian Balog, and Kjetil Nørvåg. 2012. On the modeling of entities for Ad-Hoc entity search in the web of data. In *Advances in Information Retrieval*, pages 133–145. Springer Berlin Heidelberg.

Fedor Nikolaev and Alexander Kotov. 2020. Joint word and entity embeddings for entity retrieval from a knowledge graph. In *Advances in Information Retrieval*, pages 141–155, Cham. Springer International Publishing.

Fedor Nikolaev, Alexander Kotov, and Nikita Zhiltsov. 2016. Parameterized fielded term dependence models for ad-hoc entity retrieval from knowledge graph. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 435–444, New York, NY, USA. Association for Computing Machinery.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *CoRR*, abs/1904.08375.

Carol Peters, Martin Braschler, and Paul Clough. 2012. *Multilingual information retrieval: From research to practice*. Springer Science & Business Media.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *EMNLP*.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-Yet-Effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Miriam Redi, Martin Gerlach, Isaac Johnson, Jonathan Morgan, and Leila Zia. 2021. A taxonomy of knowledge gaps for wikimedia projects (second draft). *ArXiv*.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2021. mluke: The power of entity representations in multilingual pretrained language models. *CoRR*, abs/2110.08151.

Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. 1995. Large test collection experiments on an operational, interactive system: Okapi at TREC. *Inf. Process. Manag.*, 31(3):345–360.

Paolo Rosso, Dingqi Yang, and Philippe Cudré-Mauroux. 2020. Beyond triplets: Hyper-relational knowledge graph embedding for link prediction. In *Proceedings of The Web Conference 2020*, WWW '20, page 1885–1896, New York, NY, USA. Association for Computing Machinery.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert:a hebrew large pre-trained language model to start-off your hebrew nlp application with.

Wei Shen, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. 2021. Entity linking meets deep learning: Techniques and solutions. *IEEE Transactions on Knowledge and Data Engineering*.

Chen-Tse Tsai and Dan Roth. 2016. Cross-lingual wikification using multilingual embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 589–598, San Diego, California. Association for Computational Linguistics.

Niels van der Heijden, Helen Yannakoudakis, Pushkar Mishra, and Ekaterina Shutova. 2021. Multilingual and cross-lingual document classification: A meta-learning approach. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1966–1976.

Johannes M. van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P. de Vries. 2020. Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20. ACM.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language models are open knowledge graphs. *CoRR*, abs/2010.11967.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1112–1119. AAAI Press.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S Yu. 2021. A comprehensive survey on graph neural networks. *IEEE Trans Neural Netw Learn Syst*, 32(1):4–24.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Weidi Xu, Xingyi Cheng, Kunlong Chen, and Taifeng Wang. 2020. Symmetric regularization based BERT for pair-wise semantic reasoning. In *SIGIR*, pages 1901–1904. ACM.

Puxuan Yu, Hongliang Fei, and Ping Li. 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.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Entity Retrieval Illustration

An overview of the task setup and our proposed system can be seen in Figure 3.

B Characteristics of the Various Models

In Section 3 we presented various configurations of BERTE and Table 5 compares them when used on a set of languages L. For each configuration, the table lists the number of pretrained LMs used, the languages involved in each retrieval stage, the direction of query translation (if any), and whether the system can handle languages unseen during fine-tuning.

qtBERTE$_{en}$ needs a single LM and uses only English sources in all stages, with non-English content being ignored. Because of that, queries in other languages should be translated into English. Hence, the system may be sensitive to translation errors. Recall that, in this work, we had the English version of all queries, with no need for translation.

Each BERTE$_l$ model can support a single language. Therefore, to support all languages in L, we need an array of $|L|$ BERTE$_l$ models, each of which is initialised with a different pretrained LM. For each language $l \in L$, the retriever only considers entities covered in l.

For multiBERTE variants, on the other hand, a single multilingual LM is sufficient to support all languages in L and beyond. Multilingual LMs can then be tuned only on the English dataset or on a subset of languages (L_{few}). In all cases, scores from the target language can be mixed with scores from other languages to improve the ranking, in which case query translation from l is needed. One main difference between multi$_{\text{en}}$BERTE and multi$_{\text{few}}$BERTE is in what language the training tuples are used when fine-tuning. While multi$_{\text{en}}$BERTE uses only English tuples, multi$_{\text{few}}$BERTE uses triples in several languages.

C English Results Deep Dive

To test whether the improvements from BERTE and EM-BERT are orthogonal, we linearly combine their scores (with equal weights). This hybrid retriever outperforms each of its components significantly, achieving an nDCG$_{10}$ score of 0.571, nDCG$_{100}$ of 0.634, and MAP of 0.467. While such a retriever is cumbersome and has many dependencies, it indicates that each model is complementary to each other, and combining them further increases their performance.

Another consideration to be made is about the type of queries each method excels in. Recall that `DE-v2` consists of a set of heterogeneous entity-bearing queries assembled from various benchmarking efforts. Queries are therefore categorised into four groups based on their source. Table 6 breaks down the English results of BERTE$_{en}$ by category. For three out of the four categories, BERTE$_{en}$ and EM-BERT were significantly better than the other methods. Specifically for SemSearch, which consisted of named entities, such as "Brooklyn Bridge", all methods were comparable and achieved relatively high scores. We hypothesise that this is due to the simpler nature of these queries. The simpler queries, usually only consisting of the target's name, make keyword-based retrieval methods, such as BM25, effective for most queries. Another noteworthy fact is that, like KEWER and ESim, a mixture model that combines BM25F$_{ca}$ and a neural ranker was better than its components, indicating that, despite the deep representation of entities in BERTE$_{en}$, term matching based techniques were still extremely valuable in many scenarios. Between BERTE$_{en}$ and EM-BERT, BERTE$_{en}$ had better performance for INEX-LD (IR style queries), while EM-BERT was better at QALD-2 (natural language questions).

D BERTE$_{en}$ Query Analysis

We analyse several queries to study the effectiveness of BERTE$_{en}$. For the query "What is the capital of Canada?", BERTE$_{en}$ ranks the entity for "Ottawa" in the 6th position, while KEWER and ESim leave it outside the top 10. Meanwhile, for the query "Ellis college", there is only one relevant judged entity, "Ellis University". While other methods focused on people named Ellis or on institutes with "college" in their name, BERTE$_{en}$ ranked "Ellis University" in the top 10. It shows how BERTE$_{en}$ properly leveraged the contextual similarity of "college" and "university". to rank the correct entity. In their work, Nikolaev and Kotov (2020) specifically mentioned the query "goodwill of michigan" as one where KEWER struggled with disambiguation ("Goodwill Games" vs "Goodwill Industries"). BERTE$_{en}$, however, had no problems with this query, with most top 10 results being correctly related to "Goodwill Industries".

One of the queries where BERTE$_{en}$ underper-

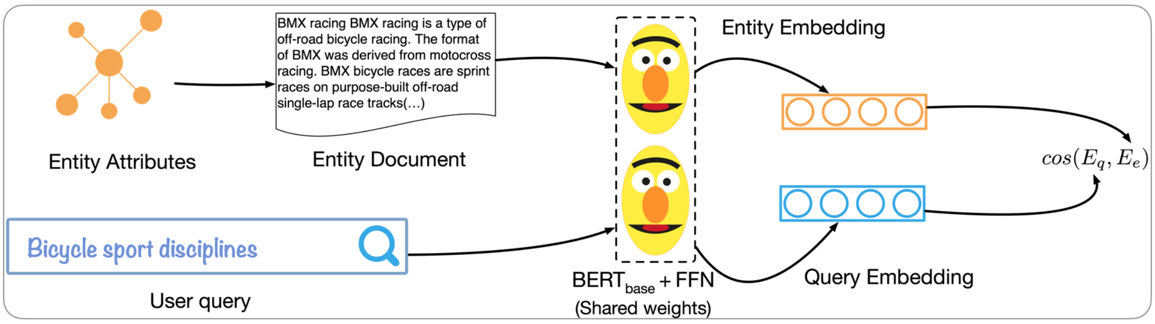
Entity Attributes
BMX racing BMX racing is a type of off-road bicycle racing. The format of BMX was derived from motocross racing. BMX bicycle races are sprint races on purpose-built off-road single-lap race tracks(...)
Entity Document
Bicycle sport disciplines
User query
BERT$_{\text{base}}$ + FFN
(Shared weights)
Entity Embedding
Query Embedding
$cos(E_q, E_e)$

Figure 3: Entity documents are generated by concatenating their literal attributes. BERTE starts with a set of candidate entities. Queries and entity documents are fed separately through the same BERT model and a fully connected layer, resulting in two vector embeddings. Final relevance estimation is computed using cosine similarity.

Table 5: Let l be the target language, L be the set of languages available to the system, and L_{mix} and L_{few} be subsets of L used for fine-tuning and mixing respectively. Here we summarise the different characteristics of the models we explore.

Model	**No. of LMs**	**First Stage**	**Fine-tuning**	**Reranking**	**Query translate**	**Handle unseen**
qtBERTE$_{en}$	1	en	en	en	$l \rightarrow$ en	$\surd$
$\{\text{BERTE}_{l^\oplus} \| l^\oplus \in L\}$	$\|L\|$	l	l	l	–	$\times$
multi$_{\text{en}}$BERTE	1	l	en	l	–	$\surd$
multi$_{\text{few}}$BERTE	1	l	L_{few}	l	–	$\surd$
BERTE$_l{}^{L_{\text{mix}}}$	$\|L\|$	$\{l\} \cup L_{\text{mix}}$	$\{l\} \cup L_{\text{mix}}$	$\{l\} \cup L_{\text{mix}}$	$l \rightarrow L_{\text{mix}}$	$\times$
multi$_{\text{en}}$BERTE$^{L_{\text{mix}}}$	1	$\{l\} \cup L_{\text{mix}}$	en	$\{l\} \cup L_{\text{mix}}$	$l \rightarrow L_{\text{mix}}$	$\surd$
multi$_{\text{few}}$BERTE$^{L_{\text{mix}}}$	1	$\{l\} \cup L_{\text{mix}}$	L_{few}	$\{l\} \cup L_{\text{mix}}$	$l \rightarrow L_{\text{mix}}$	$\surd$

Table 6: Results by query category. The following symbols indicate statistically significant improvement over: ESim ($\star$), KEWER (†), EM-BERT ($\diamond$) , and BERTE$_{en}$ ($\circ$). Best result in each column is in boldface.

	SemSearch			INEX-LD			QALD-2			ListSearch		
Model	**nDCG$_{10}$**	**nDCG$_{100}$**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**MAP**
BM25F$_{ca}$	0.628	0.72	0.529	0.439	0.5296	0.341	0.3689	0.461	0.305	0.425	0.511	0.359
KEWER	0.661	0.733	**0.563**	0.467	0.53	0.342	0.467	0.53	0.315	0.44	0.521	0.375
ESim	0.660	0.736	0.55	0.466	0.552†	0.364†	0.39	0.483	0.326	0.452	0.535	0.386
EM-BERT	0.664	**0.744**	-	0.479	0.561†	-	**0.483**$^{\star\dagger\circ}$	**0.543***†	-	**0.544**$^{\star\dagger\circ}$	0.579*†	-
BERTE$_{en}$	**0.669**	0.734	0.557	**0.509**$^{\star\dagger\diamond}$	**0.585**$^{\star\dagger\diamond}$	**0.392***†	0.441*†	0.521*†	**0.361***†	0.499*†	**0.58***†	**0.434***†

Table 7: Reranking results using multi$_{\text{few}}$BERTE and multi$_{\text{few}}$BERTE$^{\{\text{en}\}}$ for a range of additional languages. The best result for each language and metric pair is in boldface.

	BM25			multi$_{\text{few}}$BERTE			multi$_{\text{few}}$BERTE$^{\{\text{en}\}}$		
Model	**nDCG$_{10}$**	**nDCG$_{100}$**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**MAP**	**nDCG$_{10}$**	**nDCG$_{100}$**	**MAP**
Dutch	0.223	0.265	0.184	0.26	0.305	0.215	**0.317**	**0.411**	**0.262**
German	0.208	0.25	0.171	0.254	0.296	0.208	**0.361**	**0.459**	**0.297**
Turkish	0.155	0.184	0.129	0.181	0.214	0.15	**0.254**	**0.332**	**0.212**
Portuguese	0.202	0.243	0.158	0.24	0.288	0.191	**0.336**	**0.431**	**0.279**
Farsi	0.218	0.276	0.185	0.249	0.31	0.209	**0.285**	**0.374**	**0.232**
Russian	0.177	0.214	0.138	0.205	0.246	0.16	**0.345**	**0.441**	**0.286**

Table 8: Statistics of the studied chapters with respect to DE-v2. The last two rows differ due to entities relevant to more than one query.

	English	**Spanish**	**Hebrew**	**Arabic**
Has abstract	4,641,784	1,100,382	161,769	368,330
No English	-	383,963	36,710	135,527
DE-v2 Judged	45,685	17,028	6,749	7,924
DE-v2 Relevant	16,700	7,082	2,629	3,025

formed, however, is "Madrid". Examining the results, however, shows that the poor performance can be attributed in part to the annotation step. Without context, this query is open-ended and ambiguous. However, in its top 10, BERTE$_{en}$ included 7 Madrid-based sports teams, of which only 3 were judged as relevant. For example, the entity "Real Madrid C.F.", an arguably highly relevant entity that was included in the top-10 by BERTE$_{en}$, was not judged by the annotators. On the other hand, the top 10 lists of ESim and KEWER were more diverse and better matched the annotators.

E DBpedia Language Chapters

DE-v2 consists of 467 queries, with entities drawn from the 2015-10 dump from DBPedia. Additionally, relevance assessments are provided for 49,280 query-entity pairs using a 0–2 scale, with 0 being not relevant and 2 highly relevant.

The size of Wikipedia, and thus DBpedia, varies significantly across languages. Table 8 provides statistics of DBpedia 2015 for the languages we studied. English has the largest number of entities, while Arabic and Hebrew are significantly smaller, with Spanish somewhere in the middle. In the context of DE-v2, the lower coverage results in a smaller number of judged entities. Thanks to the Wikimedia foundation's efforts (Redi et al., 2021), the gap between languages is narrowing, but many languages remain low-resourced, covering 10,000 entities or less, with tens of languages, as of 2022, with chapters even smaller than Arabic or Hebrew in 2015.

F Additional Languages

Recall that multi$_{\text{en}}$BERTE and multi$_{\text{few}}$BERTE can be used for over 100 languages without fine-tuning the models on these. This allows multi$_{\text{en}}$BERTE to be a strong baseline for many languages on the task, thanks to the domain knowledge obtained in the pretraining stage and to the cross lingual capabilities of multilingual LMs. Table 7 lists the results for six additional languages, obtained following the same methodology as our main results. multi$_{\text{few}}$BERTE$^{\{\text{en}\}}$ was consistently the best performer and its advantage over BM25 and multi$_{\text{few}}$BERTE was statistically significant.

G Missing Entities

In this work, we assumed that while information from different languages may be utilised to rank entities, only entities with coverage in the target language should be returned. There are scenarios, however, where the user would like to retrieve relevant entities even if they are covered only in other languages. Consider, for example, the query "chess world champions". Of the 93 relevant entities covered in the English chapter of DBpedia 2015, only 31 had an Arabic entity. While a user who submits this query in Arabic would typically prefer to see entities in Arabic, they may also be interested in English (or some other language) if relevant entities are unavailable in Arabic.

While the English version of DBPedia is by far the largest, Spanish, Arabic, and Hebrew still offer many entities do not have an English counterpart. For instance, the entities "Roman Ornament" and "Mais el Reem" are only present in the Arabic version. The former, arguably relevant to the query "Roman architecture" (part of DE-v2), was ranked by BERTE in the top 10 for that query. The latter, a play starring Fairuz, a famous Arab singer, demonstrates that, in some cases, entities may only be of interest to speakers of the respective language.

While we hope to explore this setup in future work, initial experiments indicate that, at least in the case of Arabic queries, allowing English entities without Arabic coverage to be returned in the first stage and blending in the English scores like in multi$_{\text{few}}$BERTE$^{\{\text{en}\}}$, can improve performance by over 30%. We would stress, however, that the way to judge non-Arabic entities is not trivial and may depend on the task.

This brings the question: Are all languages equal in terms of their relevance, or do users prefer some languages over others? We hope that truly multilingual collections will be made available to allow evaluation of this scenario.