Cross-Lingual Document Recommendations with Transformer-Based Representations: Evaluating Multilingual Models and Mapping Techniques

Tsegaye Misikir Tashu and Eduard R. Kontos

Matthia Sabatelli and Matias Valdenegro-Toro
Department of Artificial Intelligence
Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence
University of Groningen, Groningen, 9747AG
t.m.tashu.rug.nl,ediraul2001@gmail.com
m.sabatelli@rug.nl, m.a.valdenegro.toro@rug.nl

Abstract

Recommendation systems, for documents, have become tools for finding relevant content on the Web. However, these systems have limitations when it comes to recommending documents in languages different from the query language, which means they might overlook resources in non-native languages. This research focuses on representing documents across languages by using Transformer Leveraged Document Representations (TLDRs) that are mapped to a cross-lingual domain. Four multilingual pre-trained transformer models (mBERT, mT5 XLM RoBERTa, ErnieM) were evaluated using three mapping methods across 20 language pairs representing combinations of five selected languages of the European Union. Metrics like Mate Retrieval Rate and Reciprocal Rank were used to measure the effectiveness of mapped TLDRs compared to non-mapped ones. The results highlight the power of cross-lingual representations achieved through pre-trained transformers and mapping approaches suggesting a promising direction for expanding beyond language connections, between two specific languages.

1 Introduction

The rapid expansion of online information from diverse sources and the growing multilingual nature of the web underscore the escalating significance of information retrieval (IR) and recommender systems (RS). Today's web is no longer limited to a single language, but is increasingly rich in multiple languages, mirroring the multilingual capacities of its global users (Steichen et al., 2014; Tashu et al., 2023). This diversity highlights the urgent need for cross-lingual recommender systems. Traditional recommender systems often prioritize content in a single language, sidelining a wealth of multilingual documents that may hold valuable insights. This gap leads to the emergence of cross-language information access, where recommender systems suggest items in different languages based on user queries (Lops et al., 2010; Narducci et al., 2016; Salamon et al., 2021).

Machine Learning and Deep Learning, which have significantly impacted language representation and processing, are pivotal to enhancing information retrieval and recommender systems, especially in the realm of document recommendation (Tashu et al., 2023; Feng et al., 2022). With these advancements, documents ranging from historical texts and scientific papers to legal ones can be recommended more accurately. However, current recommender systems falter when content is available in various languages, often recommending documents in only the query language. In multinational contexts such as the European Union, such limitations can hinder effective policy formation.

There are two main strategies to address this gap: on the one hand, one can translate the query into multiple target languages or develop a cross-lingual representation space for documents. While this can be effective, this approach is fraught with challenges, including the need for large-scale data, the computational expense of training, and potential loss in translation, especially in domains like law that require precision. On the other hand, cross-lingual representations, which focus on creating shared embedding spaces for documents across languages, are the focal point of this study (Tashu et al., 2023). By employing mapping-aligned document embeddings and comparing their similarity with the query, it offers a computationally cheaper solution without the need for extensive fine-tuning of pre-trained large language models.

The rest of the paper is organized as follows. Section 2 presents the related works. The proposed methodology is presented in section 3. Section 4 presents the experimental setting and the datasets used in this work. The experimental results will be presented in Section 5, while the results are discussed in Section 6. Finally, the conclusions

Proceedings of the 2nd SUMEval Workshop: Scaling Up Multilingual & Multi-Cultural Evaluation, pages 39–47

will be presented in section 7.

2 Related work

The work towards generating inter-lingual and multilingual representations, which can encapsulate information across multiple languages in a unified form, has gained substantial attention in recent years. This interest spans both word-level and document-level representations. Early observations, such as those introduced by (Mikolov et al., 2013), identified that word embedding spaces across languages possess structural similarities. These insights led to the development of linear mappings from one language embedding space to another, utilizing parallel vocabularies. Subsequent works (Lample et al., 2018; Smith et al., 2017; Xing et al., 2015), have aimed to refine these cross-lingual word embeddings, mainly through modifications in space alignment methods or retrieval techniques. Techniques like averaging word vectors (Litschko et al., 2018) or leveraging cross-lingual knowledge bases like Wikipedia (Potthast et al., 2008) or BabelNet (Franco-Salvador et al., 2014) have been used to learn document-level cross-lingual representation. A notable methodology in this domain is the cross-lingual semantic indexing (CL-LSI) (Deerwester et al., 1990; Saad et al., 2014), which extends the well-known latent semantic indexing (LSI) to encapsulate multiple languages through the singular value decomposition of concatenated monolingual document-term matrices.

An emerging strategy in both word-level, sentence-level and document-level research is the use of neural network architectures. One of the pioneer works in this direction was the work by (Schwenk and Douze, 2017) where they used a deep neural network to directly encode long text passages in a language-independent manner. The work by (Artetxe and Schwenk, 2019) used a multilingual auto-encoder to generate language-independent sentence embeddings. Recently, pre-trained models such as BERT (Devlin et al., 2019) have changed the landscape of cross-lingual representation research. These models have enabled the generation of sentence encoders on multilingual unlabeled corpora without the need for parallel data (Conneau et al., 2020; Feng et al., 2022; Goswami et al., 2021; Litschko et al., 2022). Concurrently, certain studies have leveraged pre-trained multilingual transformers for cross-lingual information retrieval (IR). The work by (Shi et al., 2020) combined mBERT with Google Translate in their information retrieval pipeline, while Litschko et al. (2022) utilized mBERT and XLM for the same purpose, emphasizing the need for fine-tuning for efficient and effective document-level results. Collectively, these studies underscore the potential of transformers in cross-lingual information retrieval, paving the way for alternative methodologies such as mapping over fine-tuning, as explored in the current investigation. While these approaches have shown promise, the study herein differentiates itself by presenting a methodology that uses mapping methods to create inter-lingual representations. The novelty of this work primarily lies in the use of mapping methods to align monolingual representations obtained separately for each language from pre-trained large language models, to produce inter-lingual document-level representations.

3 Methods

In this section, we will introduce the different large language models used in this study and the mapping approaches used to learn interlingual representation from the pre-trained large language models.

3.1 Transformers

Transformers, introduced by Vaswani et al. (2017) have transformed the landscape of natural language processing (NLP). Instead of relying heavily on recurrent or convolutional layers, transformers leverage multiple attention heads to weigh the significance of different parts of an input sequence differently, allowing for parallel processing and the capture of long-range dependencies in data. There exist a plethora of variations within the transformer architecture. In the following sections, we will discuss the specific variants of transformer-based large language models used in the context of this study.

3.1.1 mBERT

Multilingual BERT is an extension of the Bidirectional Encoder Representation from Transformers (BERT) that was introduced by Devlin et al. (Devlin et al., 2019). BERT stands out as a pre-trained model, having undergone training on vast volumes of unlabelled data, primarily focusing on two pre-training objectives:

- **Masked Language Modelling (MLM):** This objective requires the model to predict masked

portions of the provided input. Specifically, 15% of the training data tokens undergo masking. Of these masked tokens, 80% are substituted with the "[MASK]" placeholder, 10% are replaced with a random token, and the remaining 10% are left unaltered.

- **Next Sentence Prediction (NSP):** BERT's versatility allows it to manage tasks that involve pairs of sentences, which may or may not exhibit contextual coherence. During its training phase, BERT was supplied with sentence pairs where 50% of the pairs were contextually sequential from the training dataset, while the remaining 50% constituted random, unrelated sentences.

BERT was originally pre-trained on a strictly monolingual English corpus. Recognizing the limitations of such a unilingual approach, there emerged a demand for a model with broader linguistic capabilities. In response, the Multilingual BERT (mBERT) (Devlin et al., 2019), was conceptualized. This iteration extends the foundational principles of BERT, accommodating text from a diverse array of 104 languages.

3.1.2 mT5

Multilingual(Xue et al., 2020) Text-to-Text Transfer Transformer (mT5) is an encoder-decoder model pre-trained on 101 languages, closely based on the original T5 model from (Raffel et al., 2019). It has been pre-trained on an objective similar to MLM, called MLM span-corruption, where consecutive tokens from the input are masked from the model during pre-training.

mT5 is highly specialised for text-to-text tasks such as machine translation and text generation, however, it can also be used as an encoder model only, which was done for this project. Like BERT, the maximum amount of tokens that were used was 512, with an embedding dimensionality of 768, corresponding to the "base" version.

3.1.3 XLM-RoBERTa

The Cross-Lingual Modelling for Robustly Optimised BERT, colloquially termed XLM-RoBERTa, stands as a notable iteration of pre-trained multilingual transformers. Introduced by Conneau et al. (2019), this model is an evolution of RoBERTa (Liu et al., 2019). Diverging from conventional methodologies, XLM-RoBERTa eschews both the Next Sentence Prediction (NSP) and translation objectives, concentrating exclusively on Masked Language Modelling (MLM). The key innovation lies in refining the training procedure and extending the training duration, measures that synergistically enhance model performance. Adapted to cater to 100 languages, XLM-RoBERTa can function effectively as an encoder-only model. For the purposes of this research, the "base" variant of XLM-RoBERTa was deployed, accommodating a maximum of 512 tokens and featuring an embedding dimensionality of 768.

3.1.4 ErnieM

The Multilingual Ernie (ErnieM) (Ouyang et al., 2021) represents a distinguished pre-trained multilingual transformer. Drawing inspiration from the XLM-RoBERTa, ErnieM's hallmark feature lies in its capacity to synchronize linguistic representations across its embedded languages. This harmonization is operationalized through a cross-lingual semantic alignment, juxtaposing parallel data with its monolingual counterpart. In the spirit of achieving this, the authors put forth two pre-training objectives:

- Cross-Attention MLM (CAMLM): A strategy devised to cohesively align the semantic representation of parallel data across the entire linguistic spectrum.

- Back-Translation MLM (BTMLM): This objective embarks on aligning cross-lingual semantics with monolingual contexts. Through back-translation, it facilitates the generation of novel linguistic tokens from monolingual corpora, and subsequently acquaints the model with their multilingual semantic alignment.

Supplemented by the translation modelling language task (an initiative akin to MLM but marked by the amalgamation of sequences from an array of languages) and the Multilingual MLM (characterized by masking tokens transcending diverse languages), these objectives jointly constitute the pre-training paradigm of ErnieM. Maintaining consistency, this study harnesses the "base" version of ErnieM, with a stipulated threshold of 512 tokens and an embedding dimensionality set at 768.

The models selected for this investigation inherently embrace a multilingual ethos, underpinned by two pivotal reasons: Firstly, the monolingual iterations of these models have not ubiquitously

undergone training across the selected quintet of languages earmarked for this research. More critically, the inherent overlap in the models' embedding space across languages posits a fertile ground to evaluate the potential of leveraging ready-made multilingual models sans the requisite of supplementary mapping or precision-tuning. To draw an illustrative parallel, juxtaposing disparate models of analogous frameworks, each tailored to individual languages (e.g., BERT vis-à-vis its Gallic analogue), might yield embeddings that, owing to divergent training trajectories, manifest disparities too profound to be semantically reconciled.

3.2 Mapping approaches

Given two monolingual document collections, $D_x = \{d_{x,1}, \ldots, d_{x,n}\}$ in language x and $D_y = \{d_{y,1}, \ldots, d_{y,n}\}$ in language y . To embark on a nuanced analysis of these documents, it is imperative first to learn or extract the embedding for each document. To achieve this, we employ the pre-trained large language models introduced in section 3 subsection 3.1. Notwithstanding, it's worth noting that any representation learning algorithm that embeds the document sets D_x and D_y into vectors within the space $\mathbb{R}^k$ can be used.

From the language models, we obtain sets of vectors, respectively, defined as $C_x = \{\hat{d}_{x,1}, \ldots, \hat{d}_{x,n}\} \subset \mathbb{R}^k$ and $C_y = \{\hat{d}_{y,1}, \ldots, \hat{d}_{y,n}\} \subset \mathbb{R}^k$. Conceptually, C_x and C_y can be interpreted as "Conceptual Vector Spaces", encapsulating broader linguistic and thematic abstractions inherent to the original documents. Nevertheless, a salient point to recognize is that even if vectors within C_x and C_y encapsulate analogous concepts transversal to languages, the representation schema might vary. Consequently, a mere direct juxtaposition of $\hat{d}_{x,k}$ and $\hat{d}_{y,k}$might not manifest the underlying content congruencies.

All the mapping methods used in this study are adopted from the works of Tashu et al.(Tashu et al., 2023). In the upcoming section, we will present a summary of three different mappings where more details on each of the methods can be found in (Lenz et al., 2021; Tashu et al., 2023).

3.2.1 Linear concept approximation (LCA)

The motivation is to directly embed the test documents into the space spanned by the training documents in the semantic space using linear least squares (Salamon et al., 2021). This is based on the assumption that the vector space spanned by the parallel training documents is the same in their respective language. Therefore, the coordinates of the test documents in that span would be a good language-independent representation of these documents. Using the representation obtained from the large language models presented in section 3, we can derive low-dimensional representations of each document within $\mathbb{R}^k$. Multiple documents can be concatenated into matrices. If there are n documents available in both languages, we can create the representation/concept matrices $C_x = X^T \in \mathbb{R}^{n \times k}$ and $C_y = Y^T \in \mathbb{R}^{n \times k}$ in which every column is a concept in its respective language.

3.2.2 Linear Concept Compression (LCC)

The motivation behind LCC is to find mappings into an inter-lingual space, EC_x, C_y, such that the comparison of $C_x(\hat{d}_{x,k}), C_y(\hat{d}_{y,k})$, provides a measure of content similarity. For two monolingual representations, we want to find their inter-lingual representations, which encode the same information as the different monolingual spaces do. More precisely, for a given document d and its representations in each respective language, $\hat{d}_{x,k}$ and $\hat{d}_{y,k}$, we want to find mappings C_x and C_y, respectively, such that $C_x(\hat{d}_{x,k}) = C_y(\hat{d}_{y,k})$ and the information of $\hat{d}_{x,k}$ and $\hat{d}_{y,k}$ is preserved. The intuition is to train an Encoder-Decoder approach. The purpose of the Encoder is to encode monolingual representations in a language-independent space. The purpose of the Decoder is to reconstruct the monolingual representations of multiple languages from that encoding(Lenz, 2021).

3.2.3 Neural Concept Approximation (NCA)

In contrast to conventional approaches where mappings are directly derived from given vectors, C_x and C_y, the proposed methodology leverages a Neural Network to approximate these vectors. Specifically, a Feed Forward Neural Network (FFNN). Two distinct models were trained: one mapping from the source language to the target language, and the other in the reverse direction (Tashu et al., 2023).

Both models were defined in the same manner: 1 layer of 500 neurons, using the Exponential Linear Unit (ELU), with the Huber objective function, for a maximum of 250 epochs with the implementation of early stopping and a learning rate of $5 \cdot 10^{-4}$. The network's architecture consists of 3 total layers, one

input layer with dimensionality d (the dimension of a given document), followed by the hidden layer (with dimensionality $d \times 500$), and the output layer with dimensionality $500 \times d$.

4 Experiment

4.1 Data

The JRC-Acquis corpus (Steinberger et al., 2006) was used for this project because of its characteristics. It is a publicly available, sentence-aligned corpus consisting of the 22 official languages of the European Union (EU), containing legal documents pertaining to EU matters from 1958 to 2006. Since this study dealt with language pairs, only five languages were used, those being English, Romanian, Dutch, German, and French, for a total of 20 ordered pairs (i.e. English $\rightarrow$ French and French $\rightarrow$ English are treated as a different pair). Since the documents for each language were not aligned, it was necessary to perform a secondary alignment for the five chosen languages such that documents were shared across the subset, resulting in $6,538$ unique documents. There were also some issues at the character level of some non-English documents from the initial dataset. For instseveralber of French documents presented corrupted letters, meaning that letters with diacritics were instead displayed in XML format (e.g. "é" displayed as "%eacute"). A preprocessing step was as such introduced to replace these corrupted variants with their original form and to remove any additional white space from the documents. The documents, at the same time, were converted from XML to a standard string format to be used by the models. In this study, 60% was used for the training set, 20% for the validation set and 20% for the test set.

4.2 Embeddings

It is necessary to represent the documents in a continuous manner to be able to apply any mapping approach. This was achieved by passing all documents, in each language, through the tokenizer and model modules of the previously discussed transformer models.

An input text undergoes several processing steps while passing through the tokenizer: it is truncated or padded to the maximum length allowed by the models ($N = 512$ tokens), after which the tokens are converted to internal ID representations stored in the vocabulary of the model, and for which the attention mask is computed. The latter part allows the model to look only at the relevant tokens in the sequence, ignoring padding tokens. Since this study only deals with the embeddings of the models and not their decoded outputs, the final hidden state from the encoder part of the models is extracted. The model computed the embedding for each token, and as such, documents are now represented as 512×768 matrices, while it is necessary to obtain a vector of size 768. This was solved by performing a global pooling operation on all of the outputted states, where tokens that were not ignored by the attention mask were averaged together. As such, documents are now represented by vectors with dimensionality 768, to be used in the following section.

4.3 Evaluation metrics

Two evaluation metrics were used to compute the performance of the mapping approaches:

- Mate Retrieval Rate: the retrieval rate of the most symmetric document; this metric evaluates how similar two documents are - the query and retrieved document. If the retrieved document is the same as the query document, that is called a mate retrieval. It is defined as:

$$MR(d) = \arg\max \mathbf{S}_d \cdot \mathbf{T}_d^T$$
$$S(d, d') = \begin{cases} 1 & d = d' \\ 0 & d \neq d' \end{cases} \quad (1)$$

  where S is the similarity between 2 documents d and d', and MR is the mate retrieval for a given document d in the source S and target language T. It can be said that a mate retrieval is successful if d and d' are the same. The equations in 1 can be combined to compute the mate retrieval rate for all documents (D), as seen in equation 2:

$$\text{RetrievalRate} = \frac{1}{|D|} \sum_{d=1}^{|D|} S(d, MR(d)) \quad (2)$$

- Mean Reciprocal Rank: this represents how high-ranked documents are, based on a similarity measure. This has been achieved using cosine similarity, defined below:

$$C(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \cdot \|d_2\|} \quad (3)$$

where the numerator represents the inner product of the vector representations of documents d_1 and d_2, and the denominator is the magnitude product of the two vectors. If the two documents are similar to each other, their cosine will be closer to 1 and will be closer to -1 if they are not similar. This equation can be used to obtain the cosine matrix similarity of all documents.

Furthermore, the rank r of a document can be defined as its cosine similarity compared to other documents obtained from the matrix cosine similarity. If it is most similar to itself in the target language, then its rank will be 1. Finally, these components can be combined to form the mean reciprocal rank:

$$\text{ReciprocalRank} = \frac{1}{|D|}\sum_{d=1}^{|D|}\frac{1}{r_d} \quad (4)$$

5 Results

The performance of the mapped (or not) embeddings was measured using the evaluation metrics defined in the previous section. Due to the large number of results that were obtained (640 total results across four transformer models, three mapping methods and no mapping, for 20 language pairs, for each evaluation metric), the final results have been averaged across models and language pairs. As such, Figures 1 and 2 only present their average evaluation metric for all dimensions. Both figures showcase significant results when comparing mapped and non-mapped embeddings. However, there is also a significant difference between embeddings mapped using NCA and embeddings mapped with the other methods.

The best mapping method across both evaluation metrics was LCA (Retrieval Rate $= 0.937$, Reciprocal Rank $= 0.958$), while the worst mapping method was NCA (Retrieval Rate $= 0.609$, Reciprocal Rank $= 0.696$). Still, all methods performed significantly better than the non-mapped embeddings (Retrieval Rate $= 0.201$, Reciprocal Rank $= 0.279$). Table 1 presents the results across all language pairs for both metrics, broken down for each transformer model and mapping method, and additionally the results obtained by Tashu et al. (2023). Using the same mapping approaches, mBERT embeddings mapped using LCA outperform all other models and mapping combinations, including those from the mentioned paper, across both metrics (RetrievalRate $= 0.963$, ReciprocalRank $= 0.975$).

Model	Mapping	MRRank	MRtRate
mBERT	None	0.2	0.115
	LCA*	**0.975**	**0.963**
	LCC	0.973	0.959
	NCA	0.84	0.781
mT5	None	0.466	0.37
	LCA*	**0.947**	**0.922**
	LCC	0.936	0.907
	NCA	0.814	0.756
XLM-RoBERTa	None	0.114	0.057
	LCA	0.948	0.925
	LCC*	**0.951**	**0.928**
	NCA	0.617	0.499
ErnieM	None	0.443	0.355
	LCA*	**0.965**	**0.949**
	LCC	0.962	0.946
	NCA	0.742	0.67

Table 1: Mean Reciprocal Rank (MRRank) and Mate Retrieval Rate (MRtRate).

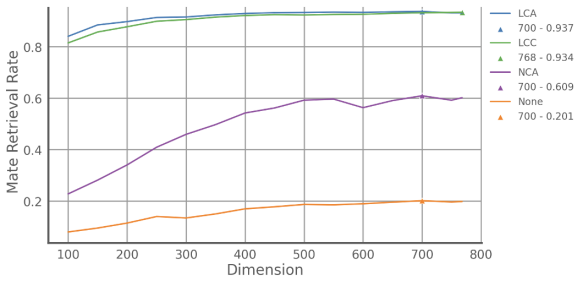

Figure 1: Line plot of the average Mate Retrieval Rate across dimensions for all language pairs and models, using LCA, LCC, NCA, and no mapping.

6 Discussion

From our results, it becomes evident that Transformer Leveraged embeddings combined with mapping methods markedly outperform non-mapped embeddings across all models, as delineated in Table 1. These Leveraged embeddings, in all instances, show significant superiority compared to the non-mapped variants. This underscores that employing an off-the-shelf model devoid of enhancements (e.g., fine-tuning, mapping) results in subpar outcomes, irrespective of the model's type. Figures 1 and 2 further substantiate this, demonstrating that mapped embeddings consistently outpace their non-mapped counterparts across all metrics. Within this context, the NCA mapping method displayed

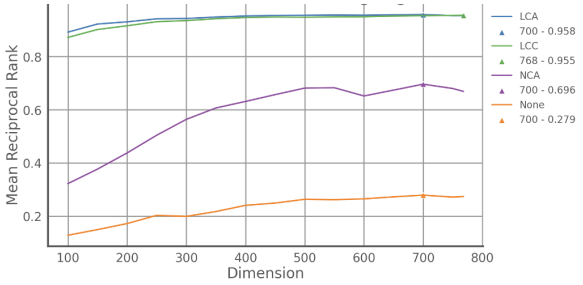

Figure 2: Line plot of the average Mean Reciprocal Rank across dimensions for all language pairs and models, using LCA, LCC, NCA, and no mapping.

the favourableorable performance, overshadowing only the non-mapped embeddings. This could be attributable to the network's architectural design, potentially falling short in capturing the nuanced similarities between documents to establish an effective mapping.

An examination of Table 1 reveals mBERT's dominance over other transformer models across all mapping strategies. Notably, when paired with LCA and LCC-mapped embeddings, mBERT eclipsed all other embedding and mapping combinations as referenced by (Tashu et al., 2023). This superior performance may be credited to the extensive data mBERT trained on, complemented by its pre-training tasks.

Interestingly, both ErnieM and mT5, when aligned with non-mapped embeddings, showcased better performance than their transformer counterparts under identical conditions. The underlying reason might be traced back to the distinctive training data and methodologies employed by these models. In contrast to mBERT and XLM-RoBERT, which utilize MLM (and additionally NSP for mBERT), ErnieM incorporates a broader spectrum of pre-training objectives geared towards cross-lingual alignment. This distinction could elucidate the superior performance of its non-mapped embeddings. mT5's commendable performance can be attributed to its foundational design, being inherently an encoder-decoder model, though this project exclusively utilized its encoding facet.

In general, our study highlights the efficacy of Transformer Leveraged embeddings when synergized with mapping techniques, resulting in a noticeable performance leap over non-mapped embeddings. This aligns with the findings of (Litschko et al., 2022), which accentuate that standalone out-of-the-box models, without refinements or supplementary techniques, are generally less efficient. However, diverging from their study, our research underscores that optimal performance doesn't solely hinge on model fine-tuning. In the realm of IR, integrating mapping techniques can be equally potent in driving commendable results.

7 Conclusion

Document recommendation stands at the forefront of Information Retrieval (IR) systems. Within recommendation frameworks, it efficiently suggests pertinent documents in alignment with a user's query. In our research, we delved into the possibilities of crafting cross-lingual representations by harnessing embeddings from pre-existing multilingual transformers in conjunction with mapping strategies. Using embeddings from these pre-trained multilingual transformers allows for document representation without requiring further training or intricate processing. Nonetheless, our research illuminated that solely depending on the raw embeddings from the transformers fell short in terms of efficacy. A notable enhancement in results was witnessed when the embeddings were synergized with mapping techniques such as LCA, LCC, and NCA. The languages incorporated within our study hold considerable prominence across various linguistic tasks. Consequently, the adopted models and mapping techniques have the potential to foster efficient representations by mapping low-resource languages onto those that are more abundantly represented. It beckons further exploration into how these mapping techniques perform when applied to low-resource languages. Future research might not restrict itself to merely language pairs, as was the focus of this study, but could expand to encompass language tuples—translating from a single source language to multiple target languages. Achieving this might necessitate refining the present mapping methodologies, introducing supplementary steps, or pioneering entirely novel methods. The code of this project is publicly available on GitHub.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Marc Franco-Salvador, Paolo Rosso, and Roberto Navigli. 2014. A knowledge-based representation for cross-language document retrieval and categorization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 414–423, Gothenburg, Sweden. Association for Computational Linguistics.

Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. 2021. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*, pages 1–14.

Marc Lenz. 2021. Learning multilingual document representations.

Marc Lenz, Tsegaye Misikir Tashu, and Tomáš Horváth. 2021. Learning inter-lingual document representations via concept compression. In *Intelligent Data Engineering and Automated Learning – IDEAL 2021*, pages 268–276, Cham. Springer International Publishing.

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*, SIGIR '18, page 1253–1256, New York, NY, USA. Association for Computing Machinery.

Robert Litschko, Ivan Vulić, Simone Paolo Ponzetto, and Goran Glavaš. 2022. On cross-lingual retrieval with multilingual text encoders. *Inf. Retr.*, 25(2):149–183.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Pasquale Lops, Cataldo Musto, Fedelucio Narducci, Marco De Gemmis, Pierpaolo Basile, and Giovanni Semeraro. 2010. Mars: A MultilAnguage Recommender System. In *Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, HetRec '10, page 24–31, New York, NY, USA. ACM.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Fedelucio Narducci, Pierpaolo Basile, Cataldo Musto, Pasquale Lops, Annalina Caputo, Marco de Gemmis, Leo Iaquinta, and Giovanni Semeraro. 2016. Concept-based item representations for a cross-lingual content-based recommendation process. *Information Sciences*, 374:15–31.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 27–38, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530, Berlin, Heidelberg. Springer Berlin Heidelberg.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Motaz Saad, David Langlois, and Kamel Smaïli. 2014. Cross-lingual semantic similarity measure for comparable articles. In *International Conference on Natural Language Processing*, pages 105–115. Springer.

Vilmos Tibor Salamon, Tsegaye Misikir Tashu, and Tomáš Horváth. 2021. Linear concept approximation for multilingual document recommendation. In *Intelligent Data Engineering and Automated Learning – IDEAL 2021: 22nd International Conference, IDEAL 2021, Manchester, UK, November 25–27, 2021, Proceedings*, page 147–156, Berlin, Heidelberg. Springer-Verlag.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167, Vancouver, Canada. Association for Computational Linguistics.

Peng Shi, He Bai, and Jimmy Lin. 2020. Cross-Lingual Training of Neural Models for Document Ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2768–2773, Online. Association for Computational Linguistics.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv:1702.03859*.

Ben Steichen, M. Rami Ghorab, Alexander O'Connor, Séamus Lawless, and Vincent Wade. 2014. Towards personalized multilingual information access - exploring the browsing and search behavior of multilingual users. In *User Modeling, Adaptation, and Personalization*, pages 435–446, Cham. Springer International Publishing.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *CoRR*, abs/cs/0609058.

Tsegaye Misikir Tashu, Marc Lenz, and Tomáš Horváth. 2023. NCC: Neural concept compression for multilingual document recommendation. *Applied Soft Computing*, 142:110348.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1006–1011, Denver, Colorado. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *CoRR*, abs/2010.11934.