Dense Information Retrieval on a Latin digital library via LaBSE and LatinBERT embeddings

Federico Andrea Galatolo[1][a], Gabriele Martino [1][b], Mario G.C.A. Cimino [1][c], Chiara Ombretta Tommasi [2]

[1]*Dept. Information Engineering, University of Pisa, 56122, Pisa, Italy*

[2]*Dept. Civilisations and Forms of Knowledge, University of Pisa, 56126 Pisa, Italy*

federico.galatolo@ing.unipi.it, g.martino8@studenti.unipi.it, mario.cimino@unipi.it, chiara.tommasi@unipi.it

Keywords: Digital Library, Information retrieval, Transformer, BERT, Latin

Abstract: Dense Information Retrieval (DIR) has recently gained attention due to the advances in deep learning-based word embedding. In particular, for historical languages such as Latin, a DIR task is appropriate although challenging, due to: (i) the complexity of managing searches using traditional Natural Language Processing (NLP); (ii) the availability of fewer resources with respect to modern languages; (iii) the large variation in usage among different eras. In this research, pre-trained transformer models are used as features extractors, to carry out a search on a Latin Digital Library. The system computes embeddings of sentences using state-of-the-art models, i.e., Latin BERT and LaBSE, and uses cosine distance to retrieve the most similar sentences. The paper delineates the system development and summarizes an evaluation of its performance using a quantitative metric based on expert's per-query documents ranking. The proposed design is suitable for other historical languages. Early results show the higher potential of the LabSE model, encouraging further comparative research. To foster further development, the data and source code have been publicly released.

1 INTRODUCTION

Information Retrieval (IR) systems have become an essential component of modern information management. These systems are designed to retrieve relevant information from large collections of documents in response to user queries. In particular, Dense IR (DIR) approaches, which are based on deep learning technology, are increasingly used in various domains, to quickly and efficiently find relevant information from large and heterogeneous data, with respect to IR based on traditional Natural Language Processing (NLP).

Recent advances in NLP have led to the development of pre-trained transformer models, such as BERT and LaBSE, that have shown impressive performance. These models are trained on massive corpora to learn rich representations of languages, and are suitable for a variety of NLP tasks.

In particular, historical digital libraries raise unique challenges. Specifically, historical texts are written in languages that are no longer in widespread use, exhibit archaic spelling and grammar, and are hard to process using traditional NLP systems. To overcome these challenges, this paper presents a DIR system for a Latin library using pre-trained transformers.

The method requires to compute the embeddings using the available models, such as Latin BERT and LaBSE, for each sentence in the documents and for the queries. To retrieve the most similar sentences for a given query, a distance such as cosine is used.

This paper delineates the development of the proposed system, which encompasses a data preprocessing pipeline, a query processing engine, and a search interface for users. To evaluate the system performance, a quantitative metric based on the per-query document ranking is provided by a Latin expert, and compared to the results achieved by the proposed system.

Experimental results show that the model based on LaBSE embeddings outperforms the one based on Latin BERT, for the purpose of retrieving pertinent information from the Latin library. To foster further development, the data and source code have been publicly released (Federico Galatolo, 2023).

[a] https://orcid.org/0000-0001-7193-3754
[b] https://orcid.org/0009-0006-3345-1045
[c] https://orcid.org/0000-0002-1031-1959

The paper is structured as follows. Section 2 covers related work. The method is discussed in Section 3. Experimental studies are covered in Section 4. Finally, Section 5 draws conclusions.

2 RELATED WORK

With the advent of machine learning, IR models have evolved from classic methods to learning-based ranking functions. One of the critical factors for designing effective IR models is how to learn text representations and model relevance matching. With the recent advancements in Pretrained Large Language Models (LLMs), such as BERT and GPT, dense representations of queries and texts can be effectively learnt in latent space, and construct a semantic matching function for relevance modeling. This approach is known as dense retrieval, as it employs dense vectors or embeddings to represent the texts (Zhao et al., 2022).

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) is a state-of-the-art language representation model that has achieved very good results on a variety of natural language processing tasks. It is a deep learning model that uses a transformer architecture to process sequences of text and generate high-quality representations. The key innovation of BERT is its use of bidirectional processing, which allows it to capture both forward and backward contextual information about a given word. This is achieved by dividing the input text into chunks of a fixed length, and then processing each chunk in both directions, from left to right and from right to left. This allows BERT to capture information about the context in which a word appears, including the words that come before and after it. In addition to bidirectional processing, BERT also uses several other techniques to improve its performance. These include the use of the following features: (i) multi-head self-attention, which allows the model to selectively focus on different parts of the input text; (ii) a masked language modeling objective, which encourages the model to predict missing words based on the context in which they appear; a next sentence prediction task, which encourages the model to understand the relationships between different sentences in a document.

David Bamman, *et al.* (Bamman and Burns, 2020) introduced Latin BERT, a contextual language model for the Latin language that was trained on a large corpus of 642.7 million words from various sources spanning the Classical era to the 21st century. The authors demonstrated the capabilities of this language-specific model through several case studies, including its use for part-of-speech tagging, where Latin BERT achieves a new state-of-the-art performance for three Universal Dependency Latin datasets. The model is also used for predicting missing text, including critical emendations, and outperforms static word embeddings for word sense disambiguation. Furthermore, the study shows that Latin BERT can be used for semantically-informed search by querying contextual nearest neighbors.

LaBSE is a multilingual sentence embedding model that is based on the BERT architecture (Feng et al., 2022). The authors systematically investigated methods for learning cross-lingual sentence embeddings by combining the best methods for learning monolingual and cross-lingual representations, including masked language modeling (MLM), translation language modeling (TLM), dual encoder translation ranking, and additive margin softmax. The authors showed that introducing a pre-trained multilingual language model dramatically reduces the amount of parallel training data required to achieve good performance. Composing the best of these methods produced a model that achieves 83.7% bi-text retrieval accuracy in over 112 languages on Tatoeba dataset, against the 65.5% accuracy achieved by previous state-of-the-art models, while performing competitively on mono-lingual transfer learning benchmarks. The authors also demonstrated the effectiveness of the LaBSE model by mining parallel data from CommonCrawl repository and using it to train competitive Neural Machine Translation (NMT) models for English-Chinese and English-German.

One recent work in language understanding that leverages contextualized features is Semantics-aware BERT (SemBERT) (Zhang et al., 2020). SemBERT incorporates explicit contextual semantics from pre-trained semantic role labeling, improving BERT's language representation capabilities. SemBERT is capable of absorbing contextual semantics without substantial task-specific changes, with a more powerful and simple design compared to BERT. It has achieved new state-of-the-art results in various machine reading comprehension and natural language inference tasks.

For latin-based IR, Piroska Lendvai *et al.* fine-tuned Latin BERT for Word Sense Disambiguation on the Thesaurus Linguae Latinae (Lendvai and Wick, 2022). This work proposes to use LatinBERT to create a new dataset based on a subset of representations in the Thesaurus Linguae Latinae. The results of the study showed that the contextualized BERT representations fine-tuned on TLL data perform better than static embeddings used in a bidirectional LSTM classifier on the same dataset. Moreover, the per-lemma

BERT models achieved higher and more robust performance compared to previous results based on data from a bilingual Latin dictionary.

More recently, Zhengbao Jiang *et al.* proposed X-FACTR: a Multilingual Factual Knowledge Retrieval from Pretrained Language Models (Jiang et al., 2020). The authors proposed a benchmark of cloze-style probes for 23 typologically diverse languages to assess factual knowledge retrieval in LLMs. The study expanded probing methods from single to multiword entities and developed several decoding algorithms to generate multi-token predictions. The results of the study provided insights into how well current state-of-the-art LLMs perform on this task in languages with more or fewer available resources. The researchers further proposed a code-switching-based method to improve the ability of multilingual LLMs to access knowledge, which has been verified to be effective in several benchmark languages. The benchmark data and code have been released to facilitate further research in this area.

3 THE PROPOSED METHOD

In this work, a cross-language DIR system is developed for Latin texts, in which the user can perform queries using different languages. For the evaluation, sample queries in Latin together with their English counterpart have been selected. The first step of the method is tokenization, i.e., breaking up all the documents into smaller text units. Tokenization depends on the specific language. Each sentence is then passed through the LLMs embeddings extraction. The proposed approach experiments the high specificity of Latin BERT and the high generality of LaBSE, although the two models are quite different. LaBSE starts from a pre-trained BERT model in Multi-lingual Model and Translation Language Model (TLM) combination, and retrains the model to combine sentence-level embeddings of different languages. In contrast, Latin BERT is trained in Masked Language Modeling (MLM), which is based on predicting a selected random word from a sentence. Both the output embeddings of the two models have a dimensionality of 768. In addition, the average embedding of the two models and the concatenated vector have been computed, to test how the combination of the two models would affect the retrieval performance. Each Latin document and each related sentence, together with its four extracted representative embeddings, are stored in a Lucene-base database that allows indexed research. When a query has to be submitted, the selected type embedding is computed and the search in the database is done using the *cosineSimilarity*, one of the most used similarity metric for search engines. It is well-known that Cosine similarity is more robust to the course of dimensionality. Finally, all the documents are sorted for similarity.

3.1 Performance evaluation

For the evaluation of the proposed DIR System, Q query sentences have been selected, both in Latin and in their respective English translations. Then, the first R results of the queries have been extracted. Finally, each resulting sentence has been evaluated via a graded evaluation between 1 and 5. The evaluation is done by a Latin-English expert. The evaluation of each retrieved document is based on the semantic coherence ratio between the query and the retrieved sentence. To evaluate each query, the Discounted Cumulative Gain (Järvelin and Kekäläinen, 2002) is used:

$$DCG_q = \sum_{i=1}^{|D|} \frac{grade_i}{\log_2(i+1)}$$

Where D is the total number of the resulting documents for the query q. The $grade_i$ is the graded evaluation in the defined interval for that result. The Normalized DCG is used: the DCG value is normalized with respect to the Ideal DCG, resorting all the results according to the evaluation. The resulting value is then in the interval [0, 1].

$$nDCG_q = \frac{DCG_q}{IDCG_q}$$

Where $IDCG_q$ (Ideal DCG) corresponds to the DCG calculated with all the retrieved documents sorted in descending order of grades. Finally, the average Normalized DCG is computed:

$$AnDCG = \frac{1}{|Q|} \sum_{i}^{|Q|} nDCG_i$$

Where $|Q|$ is the total number of queries. Let us note that the *nDCG* compares the actual document ranking of the query with respect to the ideal one, but only for the relevant documents. This raises a complication in defining the threshold for considering whether a document is relevant or not. Moreover, *nDCG* is not suitable to consider the overall performance of all the retrieved documents, besides their grade.

For this reason, a new performance evaluation metric has been developed, i.e. the *Penalized Normalized DCG*:

$$PnDCG_q = \frac{DCG_q}{IDCG_q} \times \frac{MaxDist - Dist}{MaxDist}$$

where *MaxDist* is the L1-Norm between the best possible ranking and the worst possible ranking for a set of D documents retrieved:

$$MaxDist = |D| \times |maxGrade - minGrade|$$

Dist is instead the L1-Norm between the best possible ranking and the ideal ranking:

$$Dist = \sum_{i=1}^{|D|} maxGrade - idealGrade_i$$

It is worth noting that, if the idealized ranking is exactly equal to the best grade (where all the retrieved documents have the maximum grade), the $PnDCG_q$ is equal to 1.0. In contrast, if all the retrieved documents have the minimum grade, the left-hand side of the product (the $nDCG_q$) is equal to 1.0, whereas the right-hand side becomes 0.0, bringing the whole evaluation metric to 0.0, reflecting a low performance. Finally, the Average Penalized Normalized DCG is calculated as follows:

$$APnDCG = \frac{1}{|Q|} \sum_{i}^{|Q|} PnDCG_{q_i}$$

4 EXPERIMENTAL STUDIES

For the purpose of the experiments, $Q = 10$ query sentences have been selected, both in Latin and in their respective English translations. The first $R = 10$ results of the queries have been extracted and graded by the Expert.

Table.1 shows the experimental results. It is worth noting that the LaBSE embeddings outperform any other extracted embeddings, even the combination LaBSE and LatinBERT. On the other hand, it seems that the LatinBERT embeddings are not suitable for this task, confirming the findings achieved by Hu (Hu et al., 2020). Specifically, Hu *et al.* discovered that the performance of such models on bitext retrieval tasks is very weak if not coupled with a sentence-level fine-tuning. Moreover, it is important to notice that searching the same sentence in English performs better than the respective in Latin. This bias could be ascribed to the training of LaBSE model (Feng et al., 2022): most of the sentences used for the bilingual training are in English, bringing the model to have a higher abstraction capability in English.

To further explain the performance, the PCA (Principal Component Analysis) of the embeddings has been computed, to achieve a dimensionality reduction to 50 dimensions. Then t-SNE (t-Distributed Stochastic Neighbor Embedding) is used to visualize the latent space on a bidimensional plot. To visualize the functioning of the search-engine, the t-SNE is computed with the usage of cosine similarity, to better represent the distance metric used by the search-engine.

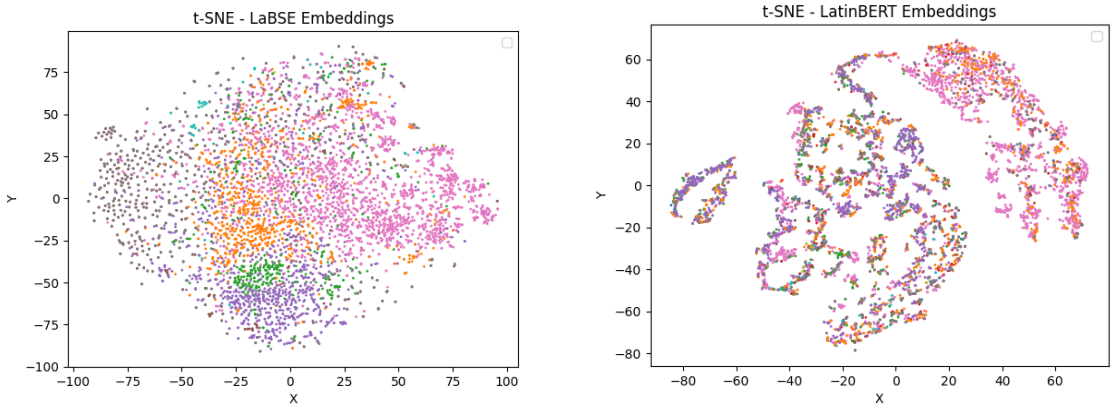
Fig.1 (a) reports the space of the LaBSE sentences embeddings for several documents. It is clear that the distribution of the LaBSE space is more regular with respect to the other space embeddings. Considering that all the documents are consistent in the topics that they treat, despite the lexical spectrum, this regularity of the space could explain the better performance achieved by the LaBSE model, as well as the lower performance of LatinBERT model, represented in Fig. 1 (b). Finally, the Mean and Concat embeddings, represented in Fig. 2 reflect a detriment of the regularity of the space.

5 CONCLUSIONS

Despite the extensive advances of Dense Information Retrieval systems (DIR), text retrieval of ancient languages, such as Latin, has to be explored, given the additional challenges. This research work illustrates the development of a DIR system, as well as its experimentation on a Latin Digital Library, carrying out multilingual queries, in Latin and English. A novel search-engine metric is also proposed, to evaluate the system performance starting from a set of graded documents. Early results show the potential of this comparative framework, encouraging further research. Specifically, it is shown that the LaBSE model outperforms the Latin Bert model, as well as that queries in English perform better than in Latin. The source code has been publicly released, along with an in-browser demonstration.

	LaBSE	LatinBERT	LaBSE-LatinBERT Mean	LaBSE-LatinBERT Concat
Latin	$\mathbf{0.33 \pm 0.06}$	0.05 ± 0.02	0.32 ± 0.06	0.32 ± 0.06
English	$\mathbf{0.52 \pm 0.05}$	NA	0.43 ± 0.06	0.43 ± 0.06

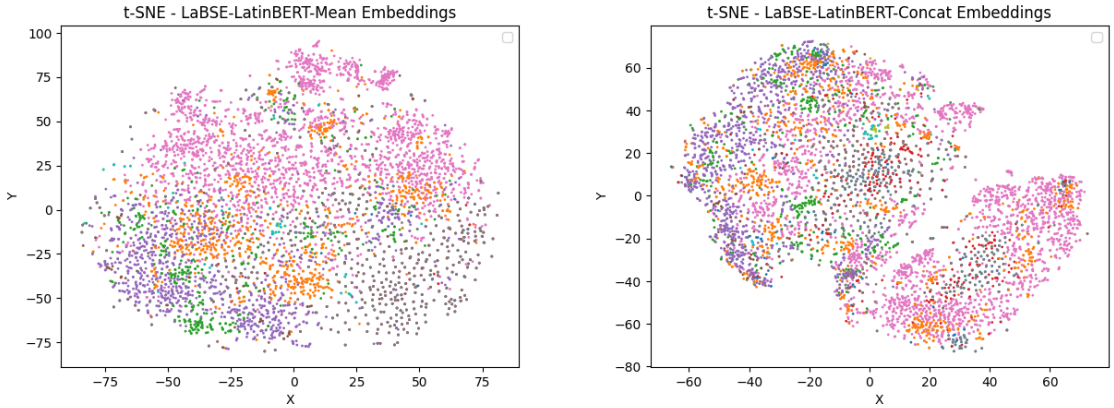
Table 1: Performance Evaluation of the proposed Dense Information Retrieval System



(a) LaBSE Embeddings Projection of documents sentences

(b) LatinBERT Embeddings Projection of documents sentences

Figure 1: Labse - LatinBERT Embeddings t-SNE



(a) LaBSE-LatinBERTMean Embeddings Projection of documents sentences

(b) LaBSE-LatinBERTConcat Embeddings Projection of documents sentences

Figure 2: Mean - Concat Embeddings t-SNE

ACKNOWLEDGEMENTS

Work partially supported by the Italian Ministry of University and Research (MUR), in the framework of: (i) the FISR 2019 Programme, under Grant No. 03602 of the project "SERICA"; (ii) the FoReLab project (Departments of Excellence); (iii) the "Reasoning" project, PRIN 2020 LS Programme, Project number 2493 04-11-2021. Research partially funded by PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme. This work has been partially carried out in the National Center for Sustainable Mobility MOST/Spoke10, funded by the Italian Ministry of University and Research in the framework of the National Recovery and Resilience Plan.

REFERENCES

Bamman, D. and Burns, P. J. (2020). Latin BERT: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Federico Galatolo, G. M. (2023). Serica-intelligent-search github. https://huggingface.co/spaces/GabMartino/serica-intelligent-search-fork, https://github.com/galatolofederico/serica-ir.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (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.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

Jiang, Z., Anastasopoulos, A., Araki, J., Ding, H., and Neubig, G. (2020). X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. *arXiv preprint arXiv:2010.06189*.

Lendvai, P. and Wick, C. (2022). Finetuning Latin BERT for word sense disambiguation on the thesaurus linguae latinae. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, pages 37–41, Taipei, Taiwan. Association for Computational Linguistics.

Zhang, Z., Wu, Y., Zhao, H., Li, Z., Zhang, S., Zhou, X., and Zhou, X. (2020). Semantics-aware BERT for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

Zhao, W. X., Liu, J., Ren, R., and Wen, J.-R. (2022). Dense text retrieval based on pretrained language models: A survey. *arXiv preprint arXiv:2211.14876*.