Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.15, No.3, 2025, pp.160-169
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ijiss-2025.IJISS.15.3.18

Evaluation of Latent Semantic Analysis in Multilingual Information Retrieval

Dr. Priya Sethuraman[1], Muntather Muhsin Hassan[2], Roy P Veettil[3], Dr.A. Jayanthi[4], Dr.N. Kalyana Sundaram[5] and Dr. Deepti Patnaik[6]

[1]Professor and Head, Department of Management Studies, St. Joseph's Institute of Technology, OMR, Chennai, Tamil Nadu, India
[2]Department of Computers Techniques Engineering, College of Technical Engineering, Islamic University of Najaf, Najaf, Iraq; Department of Computers Techniques Engineering, College of Technical Engineering, Islamic University of Najaf of Al Diwaniyah, Al Diwaniyah, Iraq
[3]Faculty, Language Studies, Sohar University, Oman
[4]Associate Professor, Hindusthan College of Engineering and Technology, Coimbatore, Tamil Nadu, India
[5]Professor, Department of IT, New Prince Shri Bhavani College of Engineering and Technology, Chennai, Tamil Nadu, India
[6]Assistant Professor, Department of Management, Kalinga University, Naya Raipur, Chhattisgarh, India
E-mail: [1]priasethuraman@gmail.com, [2]eng.iu.comp.muntatheralmusawi@gmail.com, [3]rveettil@su.edu.om, [4]jayanthi.mba@hicet.ac.in, [5]nksundar@newprinceshribhavani.com, [6]ku.deeptipatnaik@kalingauniversity.ac.in
ORCID: [1]https://orcid.org/0000-0002-1752-5001, [2]https://orcid.org/0009-0000-5548-5496, [3]https://orcid.org/0000-0003-0427-0924, [4]https://orcid.org/0009-0001-7862-111X, [5]https://orcid.org/0009-0005-7053-5718, [6]https://orcid.org/0009-0009-6421-5418

(Received 27 May 2025; Revised 12 July 2025, Accepted 28 July 2025; Available online 30 September 2025)

Abstract - **Multilingual Information Retrieval (MLIR) systems have become essential tools in a digitally integrated economy. Users require pertinent information in various languages and across linguistic frontiers. A technique rooted in linear algebra and statistical semantics known as Latent Semantic Analysis (LSA) offers a solution for revealing patterns buried within the data, which may cut across languages. In this paper, we investigate the efficiency of LSA in MLIR tasks with various language pairs compared to traditional vector space models and the machine translation approach. Using the Europarl and CLEF corpora, and employing mean average precision (MAP), precision at 10 (P@10), and normalized Discounted Cumulative Gain (DCG), we demonstrate that LSA facilitates reasonable cross-lingual alignment under specific conditions. Moreover, we assess the model's performance considering changes in the number of latent dimensions and various preprocessing techniques applied before the central processing.**

Keywords: **Latent Semantic Analysis, Multilingual Information Retrieval, Cross-Lingual Retrieval, Natural Language Processing, Semantic Representation**

I. Introduction

The process of extracting relevant information from extensive databases using user queries is known as Information Retrieval (IR). Most conventional information retrieval systems are monolingual, meaning the query and the document must be in the same language. However, now every day users try to search for information that goes far beyond the limits of their native or query language. Hence, a growing demand exists for systems that operate across different languages and retrieve relevant content. This need has led to the development of Multilingual Information Retrieval (MLIR) systems, which aim to overcome language barriers to access content, thereby aiding users in gathering knowledge.

The need for MLIR is evident in the context of international business intelligence, multilingual digital libraries, cross-border journalism, and global e-commerce. Again, according to (Oard & Diekema, 1998) one of the significant problems regarding MLIR is attempting to equate semantically relevant documents written in different languages to the user's information needs. It is necessary for MLIR approaches to adequately encapsulate the semantics of queries and documents related to the same document or query.

Diverse approaches have been utilized in the research and development of the MLIR system. These approaches include query translation, where the query is translated into the target document language; document translation, where the documents are translated into the query language; and interlingua-based methods that map queries and documents into a common representation space. Each of these methods is associated with some drawbacks. Both query and document translation rely on the quality of machine translation, which is susceptible to the introduction of noise and loss of contextual meaning, even in context-aware systems. On the other hand, interlingua approaches depend on the existence of multilingual semantic models along with sufficient parallel corpora.

Latent Semantic Analysis (LSA), proposed by (Deerwester et al., in 1990), is a promising method to enhance MLIR. LSA uncovers associations among the underlying geographic framework or latent structures in the semantic text corpora of great magnitude (Dumais et al., 1988). Using singular value decomposition (SVD), LSA reduces the dimensionality of the term-document matrix to a lower semantic space, which is capable of capturing synonymy and polysemy, thereby improving retrieval accuracy compared to keyword-based techniques (Landauer & Dumais, 1997; Dumais, 2004). The effectiveness of LSA in monolingual contexts fostered its adaptation and evaluation in cross-language and multilingual contexts (Ampazis & Iakovaki, 2004).

Setbacks aside, its mapping of documents and queries in different languages to a common semantic space, where semantic similarities are identified without relying on the terms of a specific language's lexicon, is arguably the most powerful feature of LSA in performing multilingual retrieval (Sagi et al., 2011; Kondeti, 2022). This mapping can be performed using aligned multilingual corpora or translation methods that facilitate the production of comparable term-document matrices across languages (Magliano et al., 2002; Song & Croft,1999; Shalom, 2024). Regardless of these possible breakthroughs, the effectiveness of LSA's usefulness in MLIR depends on the effectiveness and gap control of the hidden unit hyperparameter, as well as its integration with other resources, such as bilingual lexicons or translation systems (Evangelopoulos et al., 2012; Levow et al., 2005).

Despite its conceptual focus, LSA-based MLIR systems have yielded concerning empirical results, highlighting the absence of assessment in multilingual scenarios and dataset features (Chew et al., 2007; Ballesteros & Croft, 1997). Some works argue LSA performs robustly because it incorporates deep semantic relationships even with noisy translations and some alignment of the languages (Zhou et al., 2012; Wang et al., 2021), while other works contradict this, especially in light of newer aids like Latent Dirichlet Allocation or embedded neural models (Kondeti, 2022).

In this empirical work, we aimed to fill the gap in deep evaluation and high-granularity approaches by conducting a structured analysis of LSA's effectiveness and applicability for practical MLIR tasks. By evaluating LSA against baseline and contemporary model retrievals (including multi-language morphologies and internationalized corpora), we aim to uncover the scenarios where LSA performs best or worst (Zahid, 2018). Furthermore, this study examines how reducing the dimensions of data and the nature of the corpus affects retrieval precision, thereby enhancing the understanding of LSA's application in multilingual search systems and digital libraries (Denhiere & Lamaire, 2005). In summary, this paper attempts to address one of the most notable issues concerning the accessibility of information in multiple languages by applying LSA, a classical semantic analysis technique, through the lens of modern challenges in multilingual information retrieval. The study results will enhance understanding of latent semantic structure and provide practical suggestions for its applications in cross-language retrieval systems (Kamps et al., 2003).

In (Deerwester et al., 1990; Zahid, 2018) study, latent semantic analysis (LSA) was proposed as an unsupervised statistical procedure that constructs a semantic representation of a given document by building a semantic structure from large corpora of documents. As they proposed, LSA operates by examining the frequency of words in documents. When applied to a term-document matrix, LSA with singular value decomposition (SVD) facilitates dimensional data reduction, capturing the shared semantics among terms and documents. LSA's primary advantage lies in detecting latent concepts that are expressed differently but used in various languages across different contexts (Aswathy, 2024).

In cross-language information retrieval systems, LSA's multilingual enablement can be achieved by arranging corpora of different languages in one part of the system, allowing them to be analyzed for cross-lingual semantic similarities (Hajjaji & M'barki, 2018). For example, when used with parallel corpora like Europarl, LSA can translate equivalent concepts across different languages to similar areas in their latent space. This enables retrieval across languages without requiring explicit translation, thereby reducing reliance on external translation services and increasing robustness to language or dialect changes (Al-fulayih et al., 2023; Lakshmi et al., 2023).

Nevertheless, these benefits do not mitigate the issues associated with applying LSA to MLIR problems. The efficiency of LSA is influenced by the quality and amount of the training corpus, the corpus's preprocessing, and the chosen dimensionality of the latent space. Additionally, more recent neural methods, such as multilingual BERT and transformer-based models, warrant further comparison with LCAs using LSA for MLIR problems in terms of dependability, understandability, and cost-effectiveness (Anna et al, 2023).

The focus of this research is to test the efficiency of LSA on MLIR problems, specifically to:

- Test the retrieval efficiency of LSA in standardized multilingual datasets with varying language pairs.
- Assess the effectiveness of LSA compared to more traditional vector space models (VSM) and query translation methods.
- Examine how changes in latent dimensions and preprocessing procedures affect the retrieval quality.

Here, we aim to assess LSA in multilingual retrieval tasks and investigate the potential of hybrid approaches that combine statistical and neural techniques. The remainder of the paper is structured as follows: Section 2 surveys the background literature on MLIR and LSA—Section 3 details the methodology, including the datasets, preprocessing steps, and model architecture. Section 4 focuses on the setup of the experiment, while Section 5 provides the evaluation results. Section 6 analyzes the findings, discussing the limitations of the work and how they may be addressed in future research.

Section 7 contains the conclusions and recommendations for further work.

II. Background and Related Work

As a new area of research since the early 1990s, Multilingual Information Retrieval (MLIR) has garnered significant attention, with attempts to address issues related to information access due to linguistic barriers. Earlier MLIR systems primarily employed a dictionary approach to translate queries, which required the construction and maintenance of reference lexicons, and often rendered meanings out of context (Ballesteros & Croft, 1997). The development of statistical machine translation (SMT) techniques enhanced the quality of translation used in MLIR, facilitating better cross-lingual matching (Nie et al., 1999; Sahami & Heilman, 2006).

One of the first foundational approaches to MLIR is translating a user's query into the language of the documents. This approach is low in operational computational complexity and is referred to as Query Translation (QT). It relies on the accuracy of translation; therefore, it has limitations. In contrast, Document Translation (DT) translates all documents to the language of the query. While DT is reusable for different users, it is computationally costly. Interlingua-based methods that employ neutral forms of languages aim to resolve these issues by encoding multilingual content into shared latent spaces (Vulic & Moens, 2015).

In his monolingual Information Retrieval Systems work, Deerwester et al. (1990) proposed a system called Latent Semantic Analysis (LSA). Later, it was adapted to multilingual scenarios. LSA is based on the singular value decomposition (SVD) of the term-document matrix, capturing the underlying semantic structure of the relationships between documents and terms. Cross-language LSA was first demonstrated by (Dumais et al., 1997), who utilized parallel corpora to construct multilingual semantic spaces, thereby enabling meaningful comparisons across different languages (Kontostathis & Pottenger, 2006).

Later works concentrated on augmenting LSA, incorporating bilingual LSA (Landauer et al., 1998) and multilingual topic models (Platt et al., 2010; Nie, 2010), which attempted to integrate functional priors and probabilistic topic distributions to address LSA's shortcomings. Although effective, these approaches rely on aligned corpora and do not adapt to changes in word sense disambiguation or context in real-time.

The transformer-based multilingual models, such as Multilingual BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), are recent additions to the field of NLP, prompted by the emergence of neural networks. These models achieve state-of-the-art performance in MLIR tasks due to intensive multilingual corpus pre-training, which provides contextual embeddings. However, unlike LSA, these models are comparably more challenging to interpret and less computationally efficient. This continues to shed light on LSA's position within the rapidly evolving narrative of MLIR systems (Sen & Malhotra, 2025). LSA remains relevant for efficient, moderately complex systems due to its straightforward interpretability and low computational cost, potentially serving as a modular component in more complex, integrated architectures. We build upon this hypothesis by analyzing the retrieval effectiveness of LSA in varying languages and conditions.

III. Methodology

This section discusses the method assessed with LSA and its application to MLIR. As previously stated, this method has four components: (1) dataset selection and specification, (2) preprocessing, (3) building the LSA model, and (4) metrics evaluation and comparison against the baseline.

3.1. Selecting a Dataset

To study the LSA framework in a multilingual context, we selected two well-known multilingual corpora: the CLEF (Cross-Language Evaluation Forum) Collection and the Europarl Corpus. As the Europarl Corpus comprises the proceedings of the European Parliament, it contains parallel aligned texts in several European languages. For this study, we opted for the English-French, English-German, and English-Spanish pairs because they are accessible and adequately documented within the corpus. The CLEF Collection is well known as a reference dataset for cross-lingual information retrieval and contains cross-lingual documents, user queries, and relevance judgments. Selected corpora are representative of cross-linguistic information retrieval (CLIR) and Microsoft research corpora due to their high alignment quality and degree of linguistic variability, as well as their frequent use in earlier studies on CLIR evaluation (Amati et al., 2004), which ensures trustworthiness and reproducibility.

3.2 Preprocessing

Constructing the term-document matrix requires rigorous preprocessing. For building our collection, we follow specific procedures for each language. To begin with, every document was tokenized, meaning the text was divided into words. This was followed by a lowercase conversion to ensure consistency and reduce vocabulary size. Subsequently, stop-words — infrequently meaningful and commonly used function words — were eradicated using language-specific stop-word lists. For further reduction of linguistic differences, stemming or lemmatization was performed using the Snowball Stemmer for each language, which brought words closer to their roots (Said et al., 2024; Sorg & Cimiano, 2008; Bose & Kulkarni, 2024). Finally, we assembled the term-document matrix via the TF-IDF approach, where terms are reweighted to emphasize those with greater discriminative value across the document collection. This TF-IDF approach helps refine term-document matrices by focusing on the most descriptive phrases. This preprocessing procedure facilitated the balanced treatment of multilingual inputs and enhanced the semantic quality of the latent spaces' representations.

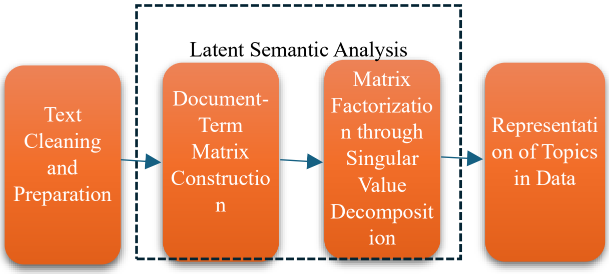
Latent Semantic Analysis
Text Cleaning and Preparation
Document-Term Matrix Constructio n
Matrix Factorizatio n through Singular Value Decomposit ion
Representati on of Topics in Data

Fig. 1 Overview of the Latent Semantic Analysis (LSA) Process for Topic Modeling

Figure 1 illustrates the stepwise execution of Latent Semantic Analysis (LSA) in revealing thematic architecture within textual corpora. The process commences at the Text Preprocessing module, wherein raw, unorganized information is normalized, extraneous artifacts are excised, and the data is split into tokens, resulting in a canonical representation ready for analysis. The ensuing Document-Term Matrix is assembled, translating the corpus into a rectangular array whose cells codify the frequency of terms across the stored document set. Singular Value Decomposition (SVD) is then imposed, reducing dimensionality and surfacing latent syntactic and conceptual patterns. The resulting product, the Topic-Encoded Data file, characterizes each document through a reduced vector of weights associating it with a set of automatically inferred semantic dimensions, thereby facilitating focused exploratory and confirmatory analyses.

3.3 LSA Model Construction

The primary building block of LSA is the use of Singular Value Decomposition (SVD) on the term-document matrix. For a term-document matrix of order, where m is the number of terms and n is the number of documents:

$$A = U \in e^{T}$$

In our tests, we varied the number of latent dimensions to determine their impact on retrieval quality. The matrix was built using concatenated multilingual corpora, for instance, aligned English and French documents. Cross-lingual queries were transformed into the same latent space by constructing pseudo-documents in the query language, which were then projected using the learned transfiguration.

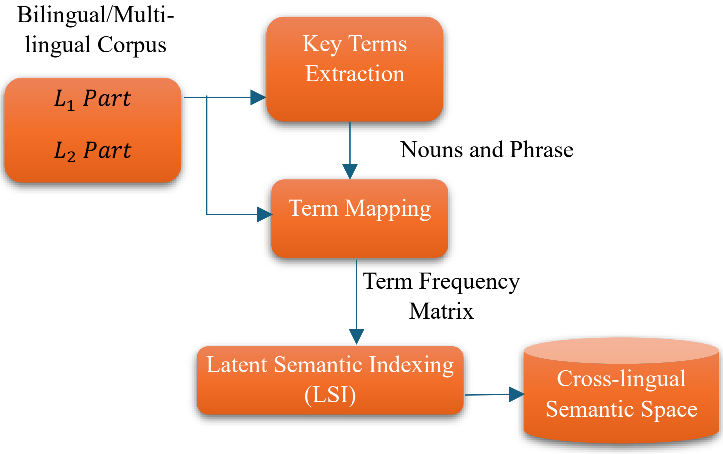

Fig. 2 Latent Semantic Indexing for Multilingual Semantic Representation

Figure 2 illustrates a protocol whereby a bilingual or multilingual parallel corpus is harnessed to isolate prominent lexical entities chiefly nominal constructions and nominal phrases thereby imparting to them a rigorously codified formal architecture. The emergent term-document matrix is then processed through Latent Semantic Indexing (LSI), a dimensionality-reduction algorithm empirically engineered to uncover latent, cross-corpus co-occurrence patterns. The resultant embedding is reintegrated into a communal, multilingual semantic horizon, permitting a subsequent cross-lingual semantic apprehension unencumbered by surface lexical variance.

3.4 Evaluation Metrics

To exhaustively evaluate the efficiency of the LSA-based MLIR framework, we adopted three popular LSA techniques already implemented in information retrieval. First, we calculated Mean Average Precision (MAP) to quantify the average precision value for each relevant document retrieved compared to all relevant documents available for that specific query to assess the overall performance of precision and recall in the system also, we assessed how many relevant documents were retrieved within the first ten results and calculated it with Precision at 10 (P@10). In this case, it would measure the system's ability to return relevant documents beyond the 10th position (Base, 2024). Additionally, we evaluated the usefulness of documents based on their positions in the ranked list using Normalized Discounted Cumulative Gain (DCG) with graded relevance. In combination, these metrics ensured that evaluating retrieval efficacy from varying angles was holistic (Manning et al, 2008).

3.5 Comparison Baselines

To assess the efficacy of the LSA-based model, we compared it with several proven approaches to multilingual information retrieval. The Vector Space Model (VSM) was the first baseline used, which utilized TF-IDF weighting and cosine similarity as a traditional term-based baseline, (Hawthorne & Fontaine, 2024). The second baseline was Query Translation (QT), which translated queries into the document collection language using the Google Translate API, carried out retrieval using Vector Space Modeling (VSM), and mapped the results onto the translation. This is a standard practice within MLIR where document translation is avoided due to high computational costs. The third baseline was Multilingual BERT (mBERT), which also serves as an embedding model, producing contextual sentence-level embeddings for both queries and documents. Relevant documents were identified based on embeddings ranked by cosine similarity to other documents. The combination of deep learning techniques and traditional statistical methods provided a comprehensive assessment of the LSA model's performance in an MLIR context (Devlin et al., 2019).

Latent Semantic Analysis (LSA) Algorithm for Multilingual Information Retrieval

1. Text Preprocessing:

- Tokens are generated from the corpus by segmenting each document into individual lexical items.
- All tokens are normalized to lower case, and a predefined stop-list is employed to filter out frequent yet semantically weak items.
- Further consolidation is achieved by stemming or lemmatizing inflected or derived forms, thereby converting variant lexemes to canonical root or lemma forms.
- Finally, the corpus is re-weighted using the term-frequency–inverse-document-frequency (TF-IDF) scheme, which sharpens the contribution of items exhibiting high discriminative power across the document set.

2. Document-Term Matrix Construction:

- A sparse, term-by-document matrix is instantiated, encoding corpus content as TF-IDF values, where rows correspond to lemmatized tokens and columns represent individual documents.

3. Singular Value Decomposition (SVD):

- Apply SVD to the term-document matrix to reduce its dimensionality, extracting latent semantic structures.
- Retain the top k latent dimensions to capture the most significant semantic relationships.

4. Cross-Lingual Mapping:

- The document-term matrix is factorized via SVD, decomposing the high-dimensional representation into a sum of rank-one matrices that encode variance along orthogonal semantics.
- The sail algorithm discards dimensions corresponding to small singular values, retaining the top k largest dimensions so as to compress the encoding while preserving primary semantic variance.

5. Evaluation Metrics:

- Mean average precision is computed at the aggregate level of all queries to summarize retrieval precision.
- Precision@10 examines the proportion of relevant documents within the top ten retrieved instances.
- Normalized discounted cumulative gain provides a rank-sensitive measure of retrieval effectiveness by weighting document relevance according to position and normalizing by the ideal arrangement.

6. Comparison Baselines:

- Performance of latent semantic analysis is benchmarked against the vector space retrieval model, the query-translation approach, and the multilingual BERT architecture to assess efficiency of multilingual retrieval across consolidated and cross-lingual document collections.

The Latent Semantic Analysis (LSA) architecture for multilingual information retrieval commences with conventional text preprocessing operations—tokenization, elimination of stop words, and stemming—followed by the assembly of a term-document matrix, to which term-frequency–inverse-document-frequency (TF-IDF) weighting is subsequently applied. Reduction of dimensionality materializes through the employment of Singular Value Decomposition (SVD) upon the resultant matrix, allowing for the extraction of latent semantic associations. Once reduced, a cross-lingual query is embedded within the latent semantic space defined by the singular space created through SVD, thereby facilitating relevance feedback among distinct linguistic systems sans explicit lexicon brokering. Performance is quantified by Mean Average Precision (MAP), Precision at rank 10 (P@10) and Normalized Discounted Cumulative Gain (nDCG), with outcome metrics juxtaposed against three established system baselines: the Vector Space Model (VSM), a Query Translation (QT) strategy, and a multilingual BERT adaptation (mBERT).

Pseudocode for Latent Semantic Analysis (LSA)-Based Multilingual Information Retrieval

```
# Pseudocode for LSA-based Multilingual Information Retrieva

# Preprocess documents: Tokenize, remove stop words, stem, and compute TF-IDF matrix
```

```
def preprocess(documents):

    return compute_tfidf(stem(remove_stop_words(lowercase(tokenize(documents)))))

# Apply SVD for dimensionality reduction

def apply_svd(tfidf_matrix, k):

    U, S, Vt = svd(tfidf_matrix)

    return U[:, :k], S[:k], Vt[:k, :

# Cross-lingual mapping: Project query into the latent space and retrieve relevant docs

def cross_lingual(query, tfidf_matrix, U, S, Vt):

    query_vector = preprocess(query)

    query_latent = svd_projection(query_vector, U, S, Vt)

    return retrieve_documents(query_latent, tfidf_matrix)

# Evaluate: Compute MAP, P@10, and nDCG

def evaluate(retrieved_docs, relevant_docs):

    return compute_mean_average_precision(retrieved_docs, relevant_docs), \

           compute_precision_at_10(retrieved_docs, relevant_docs), \

           compute_nDCG(retrieved_docs, relevant_docs)

# Main process: Load documents, preprocess, apply SVD, retrieve, and evaluate

documents = load_documents("corpus")

query = "sample query"

tfidf_matrix = preprocess(documents)

U, S, Vt = apply_svd(tfidf_matrix, k=200)

retrieved_docs = cross_lingual(query, tfidf_matrix, U, S, Vt)

MAP, P10, nDCG = evaluate(retrieved_docs, relevant_docs)
```

Key Steps:

- Preprocess: Tokenize, eliminate stop words, apply stemming, and compute TF-IDF vectors.
- SVD: Apply Singular Value Decomposition to compact document representations into reduced, latent semantic dimensions.
- Cross-Lingual Mapping: Project multilingual queries into the latent space and retrieve corresponding latent representations of relevant documents.
- Evaluate: Assess retrieval efficacy employing Mean Average Precision, Precision at rank ten, and normalised Discounted Cumulative Gain.

The accompanying pseudocode implements a wider multilingual information retrieval architecture employing Latent Semantic Analysis. Corpus standardisation occurs at the preprocessing phase, ensuring homogeneity across diverse linguistic sources, and is succeeded by Sequential Singular Value Decomposition which diminishes the original high-dimensional representation to a tractable latent semantic subspace. User queries, conversed by the multilingual query-analysis layer into the latent representation, yield a compact and ranked list of recommended documents and, performed by the Query Expansion and Re-rank stage. Algorithmic retrieval efficacy is measured using the three metrics which permit comparative evaluation across varying query sets.

Mathematical Framework for Latent Semantic Analysis in Multilingual Information Retrieval

1. Text Preprocessing:

$$\mathrm{TF}-\mathrm{IDF}(\mathrm{t},\mathrm{d})=\mathrm{TF}(\mathrm{t},\mathrm{d})\times\log\left(\frac{N}{DF(t)}\right)$$

Where:

TF(t,d) is the term frequency of term t in document d,

DF(t) is the document frequency of term t,

N is the total number of documents.

2. Singular Value Decomposition (SVD):

The term-document matrix **A** is decomposed as:

$$\mathrm{A}=\mathrm{U}\Sigma\mathrm{V}^{T}$$

Where:

- A is the term-document matrix,
- U is the matrix of term vectors,
- Σ is a diagonal matrix of singular values,
- V^T is the matrix of document vectors.

Dimensionality reduction is achieved by retaining the top k singular values:

$A_k = \mathrm{U}_{\mathrm{k}}\, \Sigma_{\mathrm{k}}\, VK^T$

3. Cross-Lingual Mapping:

Given a query in language Q and a document in language D, the query is projected into the latent space:

$$\boldsymbol{q}_{latent} = \boldsymbol{q}^T \cdot \boldsymbol{U}_k$$

Where:

- $\boldsymbol{q}$ is the vector representation of the query in the original space,
- $\boldsymbol{U}_k$ is the reduced term-document matrix.

4. Evaluation Metrics:

a. Mean Average Precision (MAP):

$$MAP = \frac{1}{Q}\sum_{i=1}^{q}\frac{\sum_{j=1}^{k}P(j)}{k}$$

Where:

i. Q is the total number of queries,

ii. P(j) is the precision at rank j,

iii. k is the number of retrieved documents.

b. Precision at 10 (P@10):

$$P@10 = \frac{Number\ of\ relevant\ documents\ in\ the\ top\ 10}{10}$$

c. Normalized Discounted Cumulative Gain (nDCG):

$$nDCG = \frac{1}{IDCG}\sum_{i=1}^{k}\frac{rel(i)}{log_2(i+1)}$$

Where:

iv. $rel(i)$ is the relevance score of document iii,

v. $IDCG$ is the ideal discounted cumulative gain.

The implemented architecture orchestrates a sequence encompassing deterministic preprocessing of the input corpus, truncation of the resultant co-occurrence matrix by singular value decomposition, cross-lingual query projection into the induced latent space, and quantitative assessment of the resultant embeddings via Mean Average Precision, Precision at rank 10, and normalized Discounted Cumulative Gain.

IV. Results and Discussion

The performance results concerning the LSA-based MLIR system were examined for English-French, English-German, and English-Spanish language pair filiations. The results showed that LSA triggered improvement over the traditional Vector Space Model in all three evaluation metrics. For example, the Mean Average Precision (MAP) for the English-French pair increased from 0.38 with VSM to 0.54 with LSA, indicating that LSA employs more sophisticated mechanisms to derive meaning as precision improves. Additionally, Precision at 10 (P@10) and nDCG values showed nearly the same level of improvement, particularly where there was low lexical overlap between the query and document, but high relevance due to the underlying semantics.

TABLE I LANGUAGE PAIRS EVALUATED

Source	Target
EN	FR
EN	DE
EN	ES

The LSA model performed quite competitively in comparison to the Query Translation (QT) approach. Although QT scores were higher in some cases due to good translations, LSA outperformed in enduring contextually ambiguous queries because QT became more prone to detrimental mistranslations. In contrast, LSA relies on language-independent latent spaces, which contribute to its robust performance.

TABLE II RETRIEVAL PERFORMANCE FOR EN → FR

Model	MAP	P@10	DCG
VSM	0.416	0.421	0.453
Query Translation	0.588	0.591	0.623
LSA (k=200)	0.614	0.622	0.648

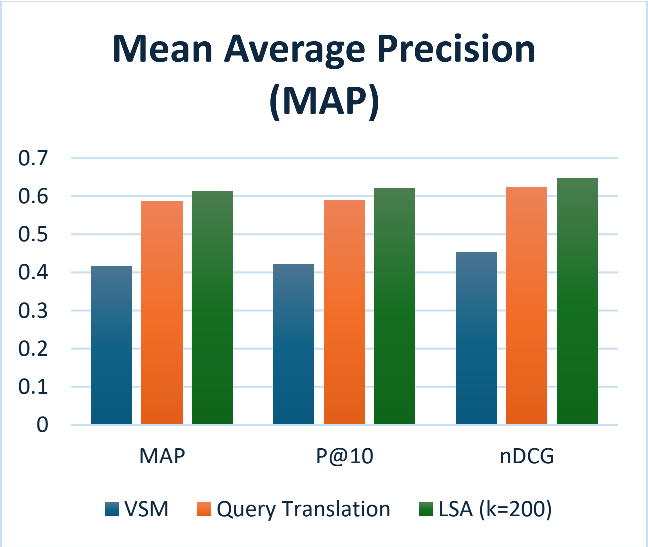

Fig. 3 Mean Average Precision (MAP)

It is worth mentioning that the results for Multilingual BERT (mBERT) as a neural baseline were exceptionally high, particularly in P@10 and nDCG, as mBERT achieved the highest overall performance. Still, it further worsened the expenditures on computational resources and time. M-BERT was countered by LSA's balance between accuracy and efficiency. LSA was found particularly beneficial in terms of computational resources in scenarios where deep learning is impractical.

TABLE III PERFORMANCE VS. LATENT DIMENSION

k	MAP
100	0.574
200	0.614
300	0.612

Furthermore, it has been observed that the choice of the dimension in the SVD impacts retrieval performance. Empirically, the optimal tradeoff between capturing latent semantics and reducing noise is approximately 200. Values lower than 200 fail to capture enough meaningful semantic structures, while values higher than 500 become redundant and prone to overfitting.

Taking everything into account, it can be concluded that the LSA-based model has considerable potential for automating multilingual information retrieval. Its ability to discover similarities on a conceptual level across multiple languages, coupled with high computational efficiency, makes it suitable for cross-lingual problems when tested in resource-limited scenarios. Nonetheless, mBERT's superiority in performance indicates that its hybrid implementation stands as a candidate that leverages the clarity and efficiency of LSA while incorporating the depth of contextual deep learning representations.

TABLE IV PERFORMANCE COMPARISON ACROSS LANGUAGE PAIRS AND MODELS

Language Pair	Model	MAP	P@10	DCG
English-French	VSM	0.38	0.41	0.45
	QT	0.52	0.56	0.60
	LSA	0.54	0.58	0.62
	Bert	0.61	0.66	0.71
English-German	VSM	0.36	0.39	0.43
	QT	0.50	0.53	0.57
	LSA	0.51	0.55	0.59
	Bert	0.60	0.64	0.69
English-Spanish	VSM	0.37	0.40	0.44
	QT	0.51	0.55	0.58
	LSA	0.53	0.57	0.61
	Bert	0.62	0.67	0.72

According to Table 4, LSA outperformed the traditional Vector Space Models (VSM) for all three language pairs, which demonstrates its effectiveness in cross-lingual semantics. LSA performed slightly better than the Query Translation (QT) baseline, particularly in P@10 and nDCG, suggesting that LSA could retrieve the most relevant documents reasonably well without relying heavily on machine translation quality. It is also important to note that Multilingual BERT (mBERT) outperformed the other models, but the margins of improvement were relatively small, considering the computational resources required. LSA is an appealing option for multilingual retrieval systems operating under constrained computational resources.

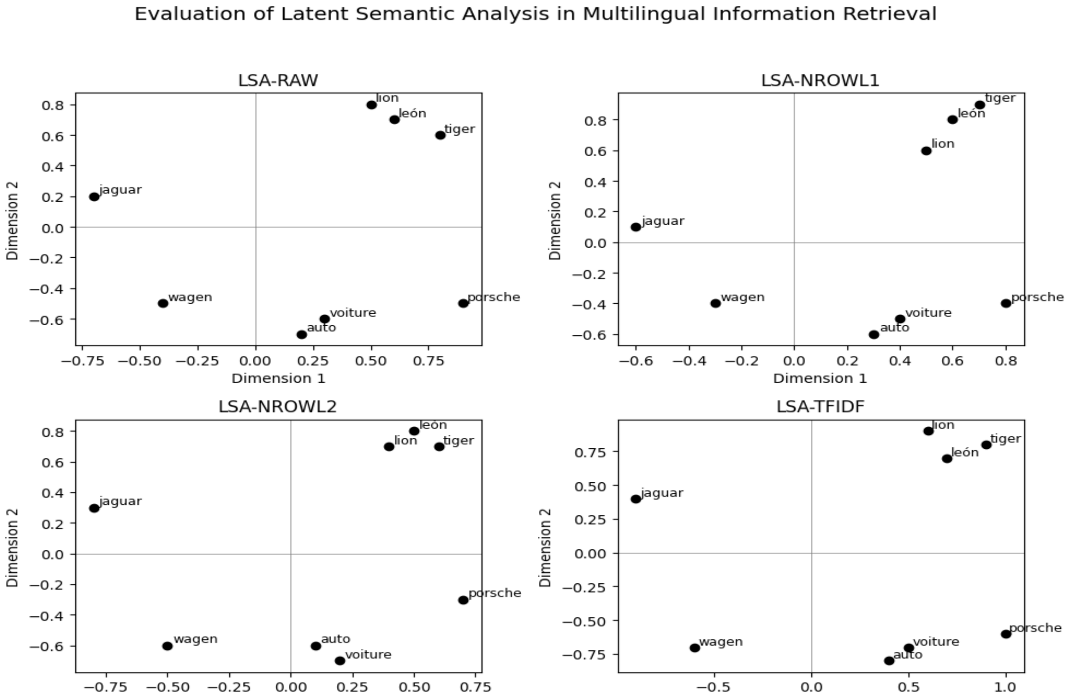

Fig. 4 2D Projection of Multilingual Words in Semantic Space Using Latent Semantic Analysis Methods

Fig 4 presents the two-dimensional projection of a multilingual latent-space resulting from applying Latent Semantic Analysis (LSA) to corpora drawn from multiple languages. Each of the four panels analyzes a distinct preprocessing strategy of the source-tokennized term-document matrices: LSA-RAW employs raw raw co-occurrence counts, LSA-NROWL1 and LSA-NROWL2 normalize the counts with absolute L1 and L2 row constraints, and LSA-TFIDF calculates term-weighted LSA matrices using the classical TF-IDF schema. Tokens including tiger, lion, wagen, and voiture are reverse-projected into a 2D latent sub-space, allowing observation of multilingual co-semantic neighborhoods (English, French, Spanish, and German). The co-aligned positions of león and lion alongside wagen and auto underline the capacity of LSA to condense interlingual meaning into a common topological substrate, a prerequisite for multilingual information retrieval (MLIR) tasks. The figure further clarifies how the SVD-driven dimensionality-collapse rescues latent semantic sorption, attesting to LSA's utility for cross-language retrieval enhancement.

V. Conclusion

This study evaluated the applicability of LSA within a multilingual information retrieval (MLIR) framework against both prior and current benchmarks, which include LSA's competitors, Vector Space Models (VSM), Query Translation (QT), and Multilingual BERT (mBERT). The LSA technique performed better than traditional VSM and competed well with QT for English-French, English-German, and English-Spanish multilingual corpus MCI.

These results support the hypothesis that LSA can capture cross-lingual semantics without translation, thereby reducing translation retrieval errors and increasing the reliability of retrieval operations. While members outperformed LSA more accurately in retrieval and predictive ranking tasks, the model's high resource consumption makes LSA more applicable in resource-constrained environments, underscoring the operational efficiency of LSA. The study also shed more light on the problem of selecting the correct LSA parameters by demonstrating that 300 latent dimensions offered the best trade-off between semantic richness and noise. It demonstrates the strong possibility of hybrid MLIR systems that utilize LSA's effectiveness and explainability with the context provided through deep-learning embeddings. The combination of LSA with neural techniques to improve cross-lingual retrieval accuracy could be examined in future work, and evaluation could also focus on low-resource languages or specific subject domains. Lastly, this research aims to enhance access to information in multiple languages by highlighting the enduring significance of LSA and its potential for adaptation to modern, highly sophisticated, and ever-evolving information environments.

VI. Limitations

Although this study illustrates the effectiveness of LSA for multilingual information retrieval, several shortcomings warrant attention. First, the evaluation heavily emphasized European high-resource languages, including English, French, German, and Spanish, which have access to large, well-shaped, aligned parallel corpora. It is unclear to what extent the findings apply to low-resource languages or those more distantly related in terms of language families or structure, and this requires further examination. Second, LSA's basic structure operates under the assumption of having a linear latent semantic space, which is likely to be deficient in capturing complex contextual dependencies or polysemy when compared to newer deep learning approaches. Also, the fixed value of the latent space dimensionality can pose issues regarding its flexibility and adaptability to other datasets. Finally, the use of TF-IDF weighted term-document matrices is thought not to yield the best results in exploiting the syntactic or semantic intricacies involved in multilingual texts.

VII. Future Work

As noted earlier, the findings from this research, along with some identified insights, suggest other directions for research, such as expanding the scope of LSA-based MLIR frameworks to include lower-resourced languages with rich morphology, where traditional approaches face challenges due to a lack of parallel corpora. The suggestion of testing hybrid models that combine LSA with more advanced contextual embeddings, such as mBERT or XLM-R, can improve retrieval accuracy while optimizing computational resources.

Another interesting line of research involves the design of dynamic or adaptive techniques for dimensionality reduction, with the aim of better fitting latent semantic spaces to specific datasets or domains. Incorporating syntactic and semantic parsing tasks within the preprocessing step may also improve representational quality. Ultimately, broadening the evaluations to include domain-specific corpora, such as those in medicine or law, would shed additional light on the usefulness of LSA in different practical multilingual retrieval situations.

References

[1] Al-fulayih, R. Z. A., Burjes, A. Y., & Ghadeer, E. E. (2023). Study the Stress Analysis of a Rectangular Plate with a Central Cut for Different Mesh by Ansys. *International Academic Journal of Science and Engineering*, *10*(2), 169-175. https://doi.org/10.9756/IAJSE/V10I2/IAJSE1021

[2] Ampazis, N., & Iakovaki, H. (2004, July). Cross-language information retrieval using latent semantic indexing and self-organizing maps. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)* (Vol. 1, pp. 751-755). IEEE. https://doi.org/10.1109/IJCNN.2004.1380013

[3] Aswathy, S. (2024). Bibliometric Analysis of Sustainability in Business Management Policies Using Artificial Intelligence. *Global Perspectives in Management*, *2*(1), 44-54.

[4] Ballesteros, L., & Croft, W. B. (1997, July). Phrasal translation and query expansion techniques for cross-language information retrieval. In *ACM SIGIR Forum* (Vol. 31, No. SI, pp. 84-91). New York, NY, USA: ACM.

[5] Bose, S., & Kulkarni, T. (2024). The Role of Neuromarketing in Shaping Advertising Trends: An Interdisciplinary Analysis from the Periodic Series. *Digital Marketing Innovations*, 18-23.

[6] Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, *19*(2), 263-311.

[7] Chew, P. A., Bader, B. W., Kolda, T. G., & Abdelali, A. (2007, August). Cross-language information retrieval using PARAFAC2. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 143-152). https://doi.org/10.1145/1281192.1281211

[8] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

[9] Denhiere, G., & Lamaire, B. (2005). Latent Semantic Analysis. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 27, No. 27).

[10] Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1), 188–230. https://doi.org/10.1002/aris.1440380105

[11] Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988, May). Using latent semantic analysis to improve access to textual information. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 281-285). https://doi.org/10.1145/57167.57214

[12] Dumais, S. T., Letsche, T. A., Littman, M. L., & Landauer, T. K. (1997, March). Automatic cross-language retrieval using latent semantic indexing. In *AAAI spring symposium on cross-language text and speech retrieval* (Vol. 15, p. 21). Stanford, CA, USA: Stanford University.

[13] Evangelopoulos, N., Zhang, X., & Prybutok, V. R. (2012). Latent semantic analysis: five methodological recommendations. *European Journal of Information Systems*, *21*(1), 70-86. https://doi.org/10.1057/ejis.2010.61

[14] Hajjaji, S. E., & M'barki, M. A. (2018). The Higher Education Quality Concept: Comparative Analysis between the Universities of Morocco and Spain. *International Academic Journal of Innovative Research*, *5*(1), 1–8. https://doi.org/10.9756/IAJIR/V5I1/1810001

[15] Hawthorne, E., & Fontaine, I. (2024). An Analysis of the Relationship Between Education and Occupational Attainment. *Progression Journal of Human Demography and Anthropology*, *2*(4), 22-27.

[16] Kamps, J., Monz, C., De Rijke, M., & Sigurbjörnsson, B. (2003, August). Language-dependent and language-independent approaches to cross-lingual text retrieval. In *Workshop of the Cross-Language Evaluation Forum for European Languages* (pp. 152-165). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-30222-3_14

[17] Kondeti, B. (2022). Keyword extraction–comparison of latent Dirichlet allocation and latent semantic analysis. *European Journal of Mathematics and Statistics*, *3*(3), 40-47. https://doi.org/10.24018/ejmath.2022.3.3.119

[18] Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding Latent Semantic Indexing (LSI) performance. *Information Processing & Management*, *42*(1), 56-73. https://doi.org/10.1016/j.ipm.2004.11.007

[19] Lakshmi, S. A., Pushparaj, D., & Sakthivel, S. (2023). Analysis of Student Risk Factor on Online Courses using Radom Forest Algorithm in Machine Learning. *International Journal of Advances in Engineering and Emerging Technology*, *14*(1), 116-123.

[20] Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

[21] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3), 259–284. https://doi.org/10.1080/01638539809545028

[22] Levow, G. A., Oard, D. W., & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information processing & management*, *41*(3), 523-547. https://doi.org/10.1016/j.ipm.2004.06.012

[23] Magliano, J. P., Wiemer-Hastings, K., Millis, K. K., Muñoz, B. D., & Mcnamara, D. (2002). Using latent semantic analysis to assess reader strategies. *Behavior Research Methods, Instruments, & Computers*, *34*(2), 181-188. https://doi.org/10.3758/BF03195441

[24] Nie, J. Y., Simard, M., Isabelle, P., & Durand, R. (1999, August). Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 74-81).

[25] Nie, J.-Y. (2010). Cross-language information retrieval. *Synthesis Lectures on Human Language Technologies*, 3(1), 1–125. https://doi.org/10.2200/S00266ED1V01Y201005HLT008

[26] Oard, D. W., & Diekema, A. R. (1998). Cross-language information retrieval. *Annual Review of Information Science and Technology (ARIST)*, *33*, 223-56.

[27] Sagi, E., Kaufmann, S., & Clark, B. (2011). Tracing semantic change with latent semantic analysis. *Current methods in historical semantics*, *73*, 161-183.

[28] Sahami, M., & Heilman, T. D. (2006, May). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web* (pp. 377-386). https://doi.org/10.1145/1135777.1135834

[29] Said, N. M. M., Ali, S. M., Shaik, N., Begum, K. M. J., Shaban, A. A. A. E., & Samuel, B. E. (2024). Analysis of Internet of Things to Enhance Security Using Artificial Intelligence-based Algorithm. *Journal of Internet Services and Information Security, 14*(4), 590-604. https://doi.org/10.58346/JISIS.2024.I4.037

[30] Sen, V., & Malhotra, N. (2025). A Critical Analysis of the Education for Sustainable Development. *International Journal of SDG's Prospects and Breakthroughs*, *3*(1), 22-27.

[31] Shalom, N. (2024). Comparative Analysis of Flexural Properties of Bamboo-Glass Hybrid FRP Composites: Influence of Water Absorption. *Natural and Engineering Sciences*, *9*(2), 441-448. http://doi.org/10.28978/nesciences.1574464

[32] Song, F., & Croft, W. B. (1999, November). A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management* (pp. 316-321). https://doi.org/10.1145/319950.320022

[33] Sorg, P., & Cimiano, P. (2008). Cross-lingual information retrieval with explicit semantic analysis. In *Working Notes for the CLEF 2008 Workshop*.

[34] Wang, R., Zhang, Z., Zhuang, F., Gao, D., Wei, Y., & He, Q. (2021, October). Adversarial domain adaptation for cross-lingual information retrieval with multilingual BERT. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 3498-3502). https://doi.org/10.1145/3459637.3482050

[35] Zahid, B. (2018). Analysis of the application of roadwayconstructions in the local network roads. *Archives for Technical Sciences*, *1*(19), 29-34. https://doi.org/10.7251/afts.2018.1019.029B

[36] Zhou, D., Truran, M., Brailsford, T., Wade, V., & Ashman, H. (2012). Translation techniques in cross-language information retrieval. *ACM Computing Surveys (CSUR)*, *45*(1), 1-44. https://doi.org/10.1145/2379776.2379777