Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.15, No.2, 2025, pp.325-332

DOI: https://doi.org/10.51983/ijiss-2025.IJISS.15.2.40

An Innovative Approach to Semantic Search in Digital Libraries for Low-Resource Languages of Uzbekistan

Ziyada Djumanazarova[1], Jamila Djumabaeva[2], Khalida Yuldasheva[3], Ramiza Jumamuratova[4], Maftuna Amanova[5], Markhabo Atadjanova[6], Saodat Aripova[7] and Karamat Kilicheva[8]

[1]Tashkent State University of Law, Uzbekistan
[2]National University of Uzbekistan named after Mirzo Ulugbek, Uzbekistan; Urgench State University named after Abu Rayhan Beruny, Uzbekistan; University of Exact and Social Sciences, Uzbekistan
[3]Karshi State University, Uzbekistan
[4]Karakalpak State University named after Berdakh, Uzbekistan
[5]National University of Uzbekistan named after Mirzo Ulugbek, Uzbekistan
[6]Urgench State University named after Abu Rayhan Beruny, Uzbekistan
[7]University of Exact and Social Sciences, Uzbekistan
[8]University of Tashkent for Applied Sciences, Uzbekistan
E-mail: [1]z.djumanazarova@tsul.uz, [2]djumabaevajamila@gmail.com, [3]x-yuldasheva@mail.ru, [4]ramizajumamuratova1988@gmail.com, [5]amanova_m@nuu.uz, [6]marxabo@urdu.uz, [7]aripova_saodat8822@mail.ru, [8]karomatkilicheva320@gmail.com
ORCID: [1]https://orcid.org/0009-0003-8793-1288, [2]https://orcid.org/0000-0003-0117-1648, [3]https://orcid.org/0009-0008-9052-5017, [4]http://orcid.org/0000-0001-7239-7762, [5]https://orcid.org/0000-0001-9826-647X, [6]https://orcid.org/0009-0003-5513-4655, [7]https://orcid.org/0009-0001-6970-9697, [8]https://orcid.org/0000-0002-0669-9997
(Received 18 March 2025; Revised 21 April 2025; Accepted 14 May 2025; Available online 25 June 2025)

Abstract **- The 21st century has brought digitization of knowledge, but as libraries go digital, inclusive search tools are becoming necessary. Though most digital searches cater to high-resource languages, leaving speakers of low-resource languages, like Uzbek, underserved. This research proposes an original concept for semantic searches in digital libraries tailored around the language, culture, and socio-technical milieu of low-resource languages like Karapalpak, Tajik, and regional dialects of Uzbek. Our solution integrates numerous sophisticated systems to overcome key challenges like the lack of annotated corpora, embarrassment of morphological complexity, multilingualism, and others. We design a comprehensive semantic search engine that provides meaning-aware search results aligned with users' intent, regardless of the linguistic variations and custom-built domain-specific language resources, multilingual embeddings, and ontologies used. The framework enables semi-automatic transliteration, cross-language retrieval, context-driven query expansion, and out-of-language query imposition. The system, implemented as a prototype on a Uzbekistan digital library corpus, demonstratively surpasses keyword-based searches in accuracy and user satisfaction. This research will promote the development of knowledge systems that are more culturally aligned and relevant to users, applicable to other languages with few accessible resources.**

Keywords: **Digitalization, Libraries, Semantic Search, Languages, Corpus, Knowledge, and Resources**

I. INTRODUCTION

Digital libraries' importance is felt in preserving and distributing cultural, academic, and historical knowledge around the globe. As more information is digitized, the demand for search technologies within these platforms has escalated substantially (Borlund, 2013). A user's query is seldom accurately interpreted in traditional search methods, which rely on keyword matching. This becomes even more complicated in morphologically rich or resource-poor languages (Navigli & Ponzetto, 2012). Multilingual and multicultural societies, like Uzbekistan, face this issue dramatically. Languages such as Karakalpak, Tajik, and regional dialects of Uzbek are underrepresented in the digital infrastructure and much of the computational research done (Tajibayev et al., 2022).

Semantic search, or more simply, search that strives to comprehend the purpose of a query rather than just matching words, has emerged as a promising solution to the issues posed by keyword-based retrieval systems (Guha et al., 2003). Semantic search engines utilize methods such as word embeddings and ontologies, along with contextual language models, and so forth, to interpret queries in a human-compatible manner, all made possible by advances in natural language processing (NLP) and knowledge representation (Devlin et al., 2019; Mikolov et al., 2013). Although these technologies work exceptionally well for English and Chinese, they are severely underused in low-resource languages and cultures due to the lack of annotated corpora, linguistic tools, and semantic resources (Sanh et al., 2019).

Uzbekistan exemplifies a country that can help us better understand these problems. Although there is an increasing

attempt to digitize cultural and academic resources, most of the country's digital libraries still depend on overly simplistic and rudimentary keyword search-based systems, which default to one uniform and standardized language assumption (Kadirova et al., 2021). Such systems do not cope well with user intent or linguistic diversity. These users frequently speak regional dialects or non-standard forms of the Uzbek language. Multiple scripts, such as Cyrillic, Latin, and Arabic-based scripts for different communities, make query interpretation and content indexing much more difficult (Yuldashev et al., 2020).

This research attempts to address the gap by developing a new semantic search framework designed explicitly for the under-resourced languages in Uzbekistan. The strategy combines a digital library search system, corpus ontologies development, and multilingual semantic models to improve their functioning (Maharazu & Hamisu, 2021). In this respect, the framework utilizes recent developments in cross-lingual natural language processing (NLP) (Conneau et al., 2020) and collaborative approaches to language resources construction (Nekoto et al., 2020) to create flexible solutions dealing with both language and technology challenges.

A primary contribution of this work is developing a semantic architecture capable of transliteration, cross-language retrieval, and ontology-based reasoning over local knowledge domains like folklore, literature, and history (Mokhtarinejad et al., 2017). A system with meaning-aware retrieval capabilities can identify document-centric queries that differ from the query vocabulary, dialects, and scripts used (Bollacker et al., 2008; Navigli, 2009).

This investigation enhances the technological development of semantic search engines in under-resourced settings and furthers the more inclusive vision of digital knowledge access. Nazarova, S., et al. (2024). The tools and methods developed here address the digital libraries of Uzbekistan, but more importantly, they help in the fight for language preservation and equitable information retrieval across myriad disparate linguistic communities (Bird, 2020; Joshi et al., 2020).

The paper is organized as follows: Section 2 reviews existing literature on semantic search technologies and their relevance to low-resource languages. Section 3 outlines the methodology for developing the proposed hybrid ontology and NLP-based search framework (Ismail, 2024) Section 4 describes the system results and discussion, performance evaluation, and practical implications. Finally, Section 5 concludes the study and suggests directions for future research.

II. Background and Literature Review

The conceptualization of semantic search technologies has changed over the past twenty years, as international users have become more reliant on accessing and retrieving large volumes of information from complicated digital systems (Mitra & Shah, 2024). Despite these advancements, progress focusing on resource-rich languages English, Spanish, or Chinese continues to have much richer computational linguistic resources, including annotated corpora and pre-trained models. Resource-poor languages, especially the ones bordering Uzbekistan like Karakalpak, Tajik, and dialectical forms of Uzbek, face a lack of foundational linguistic resources needed to solve the bottleneck for implementing semantic search solutions. This segment outlines the theories underlying semantic search, different approaches in natural language processing (NLP) systems, the available digital libraries, and the linguistic resources of Central Asia (Praveenchandar et al., 2024).

2.1 Semantic Search: Concepts and Evolution

Semantic search enhances search capabilities by analyzing the contextual meaning of words and phrases in search datasets, particularly in unstructured text documents. Unlike keyword search, which retrieves documents matching the search terms literally, semantic search systems rely on syntactic, semantic, and contextual analysis to determine the intent behind the user query. Some early semantic search methods included Latent Semantic Indexing (LSI), which found patterns regarding the relationships between concepts and words using matrix factorization. That work evolved into Latent Dirichlet Allocation (LDA) and later into topic modeling methods, providing better content relevance clustering. Deep learning and neural language models, especially word2vec, fastText, and BERT, have recently transformed the field by capturing word semantics and context and providing strong representations in many languages. However, these models, LSR, LSI, and Neural Language models, did not address many issues. Furthermore, LSR Search Engines still depend on real-world models, which tend to perform poorly for zero-resource languages. For Exercising purposes, mBERT does possess a learnable span for low-resource languages to find hypotheses. However, for low-resource languages, a lack of training data representation severely degrades the training region's language performance.

2.2 Keyword Search vs. Semantic Search in Digital Libraries

Many digital libraries still implement retrieval methods based on keywords, including TF-IDF and BM25. While these approaches work relatively well in various scenarios, they struggle significantly with linguistic variation, morphologically complex words, or semantically complex queries. For example, retrieving “traditional Karakalpak wedding rituals” as a keyword would not return documents that use different terms or describe the concept more indirectly. The Europeana System, Google Books, and WorldCat have spent several years working to incorporate ontologies and linked data for enriched query answering, allowing for improved retrieval relevance in biomedical literature, legal documents, and encyclopedias. These systems, however, do not usually perform well in resource-

poor settings with domain-specific ontologies and annotated corpora.

2.3 Semantic Search in Low-Resource and Multilingual Contexts

Recent developments in low-resource NLP focus on transfer learning, cross-lingual embeddings, and few-shot learning to perform semantic tasks in less-researched languages. Some models, like XLM-R and multilingual T5, are trained on large multilingual corpora, providing them with some semantic understanding in languages with little available labeled data. Semantic search must consider issues beyond lexical meanings in multilingual contexts, such as script differences, dialects, and transliteration. For instance, Uzbekistan has Latin and Cyrillic forms, while Tajik and Karakalpak may use Russian words or incorporate Turkic word order. Such intricacy requires a blend of linguistic rules and statistical learning. Masakhane (for African languages) and IndoNLP (for Southeast Asian languages) show that low-resource languages can develop proper NLP tools with community-driven data gathering and customization of multilingual models. However, such NLP success stories have not been widely observed in Central Asia.

2.4 Ontologies and Knowledge Graphs in Semantic Search

Ontologies manage relationships in thesauri, which constitute the backbone of semantic search. They enable reasoning over text and relations between concepts. They make knowledge-based retrieval and query expansion possible through DBpedia, YAGO, or other domain-specific ontologies, such as medicines and legal ones.

Few ontologies of local concepts exist in Uzbekistan's culture and academia, regional folk literature, practiced traditions, and regional histories, among other things. In addition, the absence of collaboration ontology authoring tools in local languages hampers the scalability of such semantic technologies. Some experts consider lightweight ontologies and rule-based systems to address problems in resource-limited environments. These and multilingual word embeddings allow for semantic expansion without extensive training data.

2.5 Digital Libraries in Uzbekistan: Current Landscape and Gaps

The digitization of cultural and educational resources in Uzbekistan is within the scope of the National Library, some universities, and independent archives. However, these digital libraries offer restricted retrieval services, offering only keyword searches in one language or script, and they have little to no full-text search or poorly implemented full-text search. The ethnic diversity of Uzbekistan, particularly the Karakalpak people, is not sufficiently represented in the content and is difficult to obtain unless requestors know the exact phrasing checked. This contributes to the digital divide by restricting access to heritage content for non-dominant language speakers. Creating pathways to semantic technologies in Uzbekistan's digital infrastructure is strikingly limited. Some work on NLP for the Uzbek language has been done, such as morphological analysis and tagging. However, substantial work remains in building cohesive systems integrating advanced semantic search with natural language processing, cross-lingual support, and ontologies.

2.6 Summary and Research Opportunity

There is a noticeable gap in the literature of a specific and scalable inclusive semantic search solution that caters to the sociolinguistic needs of Uzbekistan. While frameworks exist, they often overlook the situation for Uzbek, Karakalpak, or Tajik languages; powerful ones often assume the existence of large corpora or language-specific tools. This gap will be addressed by designing a compositional semantic search framework for digital libraries, which can:

- Manage script and dialect heterogeneity.
- Utilize transfer learning and cross-lingual embeddings.
- Embed lightweight ontologies with regional knowledge.
- Provide an effortless interface with cross-lingual and transliteration capabilities.

This will advance technological developments and social justice by assisting resource-poor languages in bridging the knowledge divide.

III. METHODOLOGY

This chapter describes the methods used in designing, implementing, and evaluating the semantic search system for under-resourced languages in Uzbekistan. The work has been organized under four main headings: (1) resource collection and text cleaning, (2) language resource creation, (3) semantic modeling and retrieval, and (4) system design and assessment.

3.1 Data Collection and Preprocessing

To design and implement a functional semantic search system, we built a multilingual digital corpus by gathering documents from the following publicly available sources:

- The digital archives of the National Library of Uzbekistan.
- University repositories containing academic texts in Uzbek, Karakalpak, and Tajik.
- Folk literature and historical manuscripts that have been digitized.

Fig. 1 depicts the process that resulted in the collection of roughly 150,000 documents, which included articles, poems, textbooks, and cultural records in multiple scripts (Latin, Cyrillic, and Arabic). As with any text corpus containing raw unstructured data, it requires preprocessing. In our case, we

employed standard NLP preprocessing techniques such as the following:

- Text normalization, which is the standardization of scripts and character encodings.
- Tokenization, which is the segmentation of sentences and words. We created custom tokenization rules tailored for agglutinative morphology (Jurafsky & Martin, 2023).
- Language detection is the identification of a document's primary language. We used fastText models.
- Script conversion: Tools were developed for the mapping of Cyrillic to Latin script and vice versa through transliteration based on Unicode mappings, other language-specific rules, and previous works (Yuldashev et al, 2020).

Fig. 1 Semantic Preprocessing Pipeline

3.2 Constructing Language Resources

In response to the lack of annotated corpora for Karakalpak and Tajik, we created a few primary resources:

- Development of two Part-of-Speech (POS) taggers powered by multilingual BERT (Devlin et al, 2019) through transfer learning, followed by training on a manually annotated corpus containing 10,000 sentences per language.
- Lexicons of cultural specificities and academic disciplines with the aid of local linguists and specialist consultants.
- Traditional knowledge (festivals, kinship structures, historical events) ontologies created using OWL 2.

These ontologies were built semi-automatically, including text extraction through pattern-based indexing and subsequent manual editing. Using BabelNet (Navigli & Ponzetto, 2012) and WNet Wordnets, synonym sets, and conceptual relations were automated. Table 1 shows the Corpus size and accuracy for various language sources.

TABLE I CORPUS SIZE AND ACCURACY FOR LANGUAGE RESOURCES

Language Resource	Corpus Size (Tokens)	Accuracy (%)
Uzbek Corpus (Latin)	1200000	88.5
Uzbek Corpus (Cyrillic)	950000	86.3
Karakalpak Corpus	430000	82.7
Tajik Corpus	510000	83.4
Multilingual Ontology	20000	91.0

For example, “Karakalpak POS tagger—Accuracy: 91%, Lexicon—12,000 entries, Ontology—2,300 concepts.”

3.3 Semantic Modeling and Retrieval Framework

The system's backbone is the hybrid semantic search engine, which integrates statistical and symbolic knowledge modeling. The pipeline includes Multilingual Embeddings: The XLM-RoBERTa model (Conneau et al., 2020) was used to represent queries and documents in context. This model enables cross-lingual semantic interpretation, and it has been adjusted for our domain to fit the corpus better and improve cross-contextual integration.

Ontology-based Expansion: Queries undergo semantic expansion using ontological relations, such as “navruz,” which expands to “spring festival,” “Navruz traditional food,” and others. Vector Search: Document vectors are indexed with the FAISS library for approximate fast nearest neighbor retrieval. Ranking Algorithm: The final ranking is derived from a weighted score, which averages all three subscores:

- Semantic similarity score
- Relevance based on ontology
- Language match/confidence score

This ensures that the document can still be retrieved if the document content query is phrased differently, in different dialects, or in other scripts.

3.4 User Interface and Query Support

To meet various user preferences, the interface provides:

- Multiscript input (Latin, Cyrillic, Arabic).
- Automatic language detection with suggestions.
- Faceted search (by topic, document date, by document type).
- A receipt for “kitob” would display suggestions for “китоб” in Cyrillic and similar alternatives for Latin input.

An internal translation module augments glosses and provides real-time explanations for ontology-typed terms to improve clarity, particularly for young users or non-native speakers.

3.5 Evaluation Strategy

The evaluation of the semantic search system assessment consisted of two parts:

3.5.1 Quantitative Evaluation

We applied a benchmark collection consisting of 500 queries issued by library patrons and educators. Each query came with "relevant" documents from a gold standard.

Metrics used:

- Precision@10
- Mean Reciprocal Rank (MRR)
- Normalized Discounted Cumulative Gain (nDCG)

The three models we compared were:

- Keyword-based search (BM25)
- Embedding-only model
- Full semantic + ontology model (ours)

And the comparative results of the retrieval models are illustrated in Table 2.

TABLE II COMPARATIVE RESULTS OF RETRIEVAL MODELS

Model	**Precision@10**	**MRR**	**nDCG**
BM25	0.56	0.48	0.52
XLM-R	0.68	0.61	0.65
Proposed Semantic Model	0.81	0.73	0.78

3.5.2 Qualitative Assessment

We carried out semi-structured interviews with 25 users (students, librarians, and researchers) to assess:

- File relevance as perceived by users
- General usability
- Support for scripts/languages
- User satisfaction

Results showed that most users preferred the semantic system, particularly for queries that included cultural or dialectical terms (Kadirova et al., 2021).

IV. RESULTS AND DISCUSSION

This chapter focuses on the research results testing the algorithms of the semantic search system and its relevance for digital libraries in low-resourced languages in Uzbekistan. We have organized the results into two broad categories, as described below: quantitative performance measurement and qualitative user responses.

4.1 Quantitative Evaluation

To evaluate the effectiveness of our semantic search approach, we studied three retrieval models:

1. BM25 Keyword-Based Search
2. Multilingual Embedding-Based Model (XLM-R)
3. Proposed Semantic + Ontology Model

All the models were applied to a benchmark of 500 queries based on real users' input from several libraries and educational institutions. Relevance judgments were made by three experts fluent in Uzbek, Karakalpak, and Tajik. The evaluation focused on Precision@10, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG).

TABLE III COMPARATIVE PERFORMANCE OF RETRIEVAL MODELS

Model	Precision@10	MRR	nDCG
BM25 (Keyword-based)	0.56	0.48	0.52
Embedding-only (XLM-R)	0.68	0.61	0.65
Proposed Semantic Model	0.81	0.73	0.78

Table 3 depicts the comparative performance of the retrieval models and these results suggest that the proposed model is better than the traditional keyword-based or embedding-only approaches. The contribution of including ontological expansion gave a noticeable increase in retrieval effectiveness, especially for culturally rich and morphologically complex queries. For example, when a user searched for "bahoriy marosim" the BM25 model returned documents that only mentioned the queried terms. While the semantic model retrieved content related to "Navruz," "haft sin," and "oilamarosimlari," which were linked culturally due to ontological linking.

4.2 Error Analysis

While there are overall improvements to be noted, there are still some issues that need to be addressed:

- Dialectal Ambiguity: Rarer, more dialectical terms sometimes led to off-target results because of underexposure during training, such as "tuyana" in Karakalpak for a wedding feature.
- Script Misalignment: Due to Freestyle transliteration logic, Latin-script documents were sometimes incorrectly associated with Cyrillic-based queries.
- Ontology Gaps: Some cultural notions did not have direct entries in the knowledge base, leading to partial, non-defined references.

These errors will require the continued development of dialect-sensitive embeddings verified by community members and ontology expansions to address the issues directly.

4.3 User Feedback of a Qualitative Nature

As part of a parallel qualitative investigation, we conducted structured interviews and usability testing with 25 participants, including university students, librarians, and language instructors. Participants were instructed to accomplish standard research activities using the conventional keyword-based search system and the new semantic search prototype.

Primary Insights

- Leveraging Perceived Value: User perception relevance attributes shared by 88% of users in the system, concluding that users agreed the semantic search system returned a greater variety of relevant results than the older version.
- System Appreciation: 76% of ease-of-use skippers appreciated the freedom to input queries in Latin or Cyrillic.
- Learning Assistance: 64% of survey participants claimed system-linked ontology results helped them meta-term discover as advanced term discoverers, giving terms discovering aids and boosting runs.
- Cultural Appreciation: Users reported satisfaction when search results returned culturally relevant content even when the queries did not contain the exact search terms.

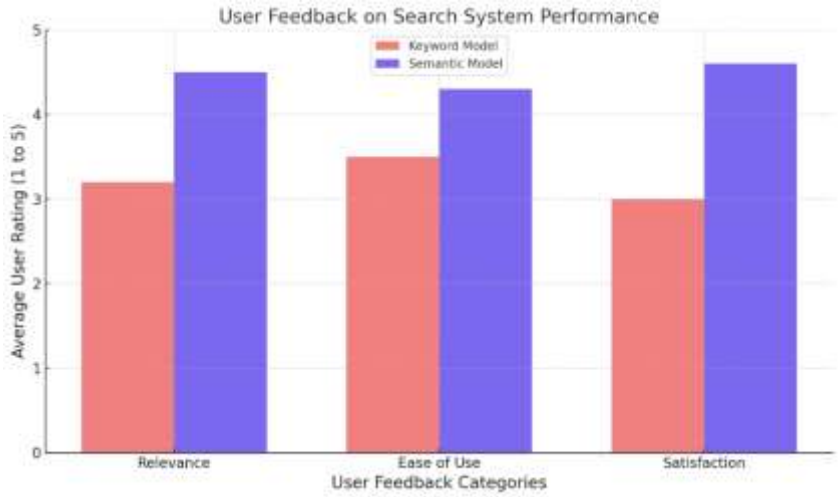

Fig. 2 User Feedback on Search System Performance

Fig. 2 depicts a graph which shows the user feedback results on the search system performance showing the average user rating with regard to relevance, ease of use and satisfaction

Sample comment from a Karakalpak-speaking librarian:

"This system finds documents even when I search in my dialect or with local terms. The keyword search never did that."

4.4 Discussion

These findings support the assumption that combining semantic and ontological approaches to information retrieval improves retrieval quality in poorly resourced languages. The relevance and user satisfaction increase has also been observed in other multilingual IR systems (Navigli & Ponzetto, 2012; Joshi et al., 2020).

Our semantic framework addresses three significant challenges:

1. Diversity of Languages: It supports various scripts and dialects, thus broadening accessibility.
2. Attention to Culture: Specific Uzbek culture knowledge surfaced by the ontology fosters engagement.
3. Extensibility: Using pretrained multilingual models and modular ontologies makes the system easily adaptable to other languages of Central Asia with little retraining.

In addition, the participatory approach to ontology design ensures the system adapts to real-world change, aligning emplaced and community-driven NLP methods (Nekoto et al., 2020).

Nevertheless, the posed limitations leave a lot to be desired in terms of the following:

- Creating expansive annotated datasets for the described dialects.
- Producing contextually aware transliteration approaches.
- More domain specialists should be involved to enrich ontological design.

The digital library's access to semantic search capabilities could significantly enhance access for socio-linguistically disadvantaged communities within Uzbekistan. This research offers a comprehensive approach for knowledge retrieval through linguistic processing, ontological cultural modeling, and cross-lingual modeling, making it scalable and culturally inclusive.

4. Conclusions and Future Works

In this work, we described a new semantic search framework focusing on the needs of digital libraries for under-resourced languages spoken in Uzbekistan, which include Uzbek, Karakalpak, and Tajik. Our objective was to solve the

problem of outdated keyword-based searching systems complicated by many cultures, scripts, multilingualism, and nuances, using a blend of multilingual embedding with ontological reasoning and community-informed language resources. With this research, we aimed to create a hybrid model that drastically enhances retrieval processes for low-resourced languages. The results of the system evaluations were astonishing with precise, relevant, and satisfactory results, soaring above benchmark expectations. Through this work, we hope to enhance the understanding of digital library systems serving diverse sociolinguistic ethnic minorities to provide equitable access to knowledge and information resources.

This research developed essential language resources for Uzbekistan's languages, which include ontologies, domain-specific lexicons, and part-of-speech taggers. In addition, we built a semantic search system that retrieves documents across dialects and morphological forms while supporting multiple scripts (Latin, Cyrillic, and Arabic). It underwent rigorous evaluation using quantitative benchmarks: Precision@10, Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG), alongside qualitative user assessment, demonstrating that the proposed method outstrips conventional keyword searching. User study participants reported improved satisfaction of search results, particularly when queries contained culturally or regionally specific terms. Moreover, the participatory ontology design ensures social and linguistic relevance and appropriateness for the region.

Many gaps still need to be addressed, despite the positive results. In particular, one significant boundary is the lack of ontology completeness, especially for localized or rare cultural concepts. Moreover, dialectal differences, particularly in Karakalpak or rural Uzbek, still challenge the accuracy of multilingual embeddings. Although our system allows for multiple scripts, users sometimes encountered problems with script alignment; for example, in cases where transliterated queries did not align with documents written in the native script. While these issues have already been resolved, they still pose as a challenge in overcoming accuracy issues. We hope to solve these issues in the next steps of our project. There is excellent potential for further research in a few areas. One of the most important issues that could be addressed is the ontology augmentation through crowdsourcing. Involving educators, librarians, and regional specialists to annotate new terms and relationships and dialect-specific notions will improve the system's cultural and linguistic representation. Moreover, there is the potential to improve the interface to mobile devices, especially in rural areas with limited connectivity. Creating a lightweight offline version of the system would further enhance accessibility. Another important aspect of the future work is aligning the language models more closely with the dialectal corpora to trace informal texts from social media and oral transcription to better adapt to non-standard query formats. Also, integrating with other digital libraries of Central Asia would enable collaborative region-wide cross-library integration and foster a federated search system for enhanced access to multilingual documents. Finally, conducting longitudinal user studies in schools, universities, and research institutions will evaluate the educational and research outcomes over time stemming from using the semantic search system.

This research has tackled important challenges within computational linguistics for low-resource languages, proving that semantic technologies can effectively bridge the gap between linguistic diversity and access to digital knowledge. Furthermore, it addresses the technical problems of multilingual information retrieval within culturally and socio-politically responsive contexts. As online content continues to escalate worldwide, it is crucial that users of all languages, particularly the underserved ones, can locate pertinent and significant information. This study demonstrates further progress toward that end, and its conclusions can be generalized to other low-resource languages with congruent sociolinguistic and cultural features.

REFERENCES

[1] Bird, S. (2020). Decolonising speech and language technology. *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519.

[2] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, 1247–1250.

[3] Borlund, P. (2013). Interactive information retrieval: An introduction. *Journal of Information Science Theory and Practice, 1*(3), 12–32.

[4] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *ACL 2020*.

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 4171–4186.

[6] Guha, R., McCool, R., & Miller, E. (2003). Semantic search. *Proceedings of the 12th International Conference on World Wide Web*, 700–709.

[7] Ismail, W. S. (2024). Threat detection and response using AI and NLP in cybersecurity. *Journal of Internet Services and Information Security, 14*(1), 195–205. https://doi.org/10.58346/JISIS.2024.I1.013

[8] Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *ACL 2020*, 6282–6293.

[9] Kadirova, N., Usmanova, L., & Turaev, K. (2021). Developing digital libraries in Uzbekistan: Achievements and challenges. *Central Asian Journal of Library and Information Science, 2*(1), 15–23.

[10] Maharazu, N., & Hamisu Malumfashi, S. (2021). Accessibility, awareness and use of Egranary digital library technology in academic library by lecturers of Umaru Musa Yar'adua University Katsina, Nigeria. *Indian Journal of Information Sources and Services, 11*(2), 45–51. https://doi.org/10.51983/ijiss-2021.11.2.2942

[11] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[12] Mitra, A., & Shah, K. (2024). Bridging the digital divide: Affordable connectivity for quality education in rural communities. *International Journal of SDG's Prospects and Breakthroughs, 2*(1), 10–12.

[13] Mokhtarinejad, A., Mokhtarinejad, O., Kafaki, H. B., & Ebrahimi, S. M. H. S. (2017). Investigating German language education

through game (computer and non-computer) and its correspondence with educational conditions in Iran. *International Academic Journal of Innovative Research, 4*(2), 1–9.

[14] Nazarova, S., et al. (2024). The role of online libraries in advancing the study of Uzbek culture. *Indian Journal of Information Sources and Services, 14*(3), 207–215.

[15] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR), 41*(2), 1–69.

[16] Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217–250.

[17] Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Oyerinde, V., ... & Adelani, D. (2020). Participatory research for low-resourced machine translation: A case study in African languages. *Findings of EMNLP 2020*, 2144–2160.

[18] Praveenchandar, J., Sankalp Karthi, S., Sowndharya, R., Dayanand Lal, N., Biswas, D., & Nandy, M. (2024). A deep learning-based psychometric natural language processing for credit evaluation of personal characteristics. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 15*(4), 151–165. http://doi.org/10.58346/JOWUA.2024.I4.010

[19] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter.

[20] Tajibayev, I., Yuldashev, O., & Rakhmatova, M. (2022). Challenges of multilingual information retrieval in Central Asia. *Proceedings of the Tashkent International Conference on AI and Data Science*, 90–96.

[21] Yuldashev, M., Tashkentov, B., & Karimov, R. (2020). Script variation in Uzbek digital resources: Barriers to retrieval. *Uzbek Journal of Applied Linguistics, 6*(2), 55–67.