INTERNATIONAL JOURNAL OF ADVANCED INNOVATIVE TECHNOLOGY IN ENGINEERING

Published by Global Advanced Research Publication House Journal

Home page: www.ijaite.co.in

Word Sense Disambiguation approach in Cross Language Information Retrieval

[1]Vivek A. Manwar, [2]Dr. A. B. Manwar, [3]Dr. Mohammad Atique

[1]Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India.

[2,3]Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India

[1]vivek.manwar007@gmail.com, [2]avinashmanwar@sgbau.ac.in, [3]mohammadatique@sgbau.ac.in

Article History

Received on: 05 Dec 2025

Revised on: 05 Jan 2026

Accepted on: 20 Jan 2026

Keywords: Information Retrieval, Cross-Language Information Retrieval, Word Sense Disambiguation, Multi-lingual Information Retrieval, Cross-lingual Information Retrieval.

E-ISSN: 2455-6491

DOI:10.65809/IJAITE/26/v11i01/001

Production and hosted by
www.insightiveinc.org

ABSTRACT

Cross-Language Information Retrieval enables users to retrieve documents written in a language different from the query language, thereby overcoming linguistic barriers in information access. However, CLIR for Indian regional languages remains challenging due to lexical ambiguity, morphological richness, and limited linguistic resources. In particular, Marathi language exhibits a high degree of polysemy and homonymy, which often leads to semantic drift during query translation and degrades retrieval performance. This paper investigates the role of Word Sense Disambiguation in improving Marathi–English CLIR and proposes a novel hybrid WSD-based CLIR framework tailored for low-resource languages. The proposed approach integrates Marathi-specific morphological analysis, knowledge-based sense inventory from Marathi WordNet, contextual semantic similarity modeling, and sense-aware query reformulation to ensure semantically faithful translation. Experimental evaluation is conducted on a Marathi query set and an English document corpus using standard information retrieval metrics. Comparative results demonstrate that the proposed hybrid WSD-based CLIR framework significantly outperforms dictionary-based, first-sense, and knowledge-based baselines, achieving superior early precision and ranking effectiveness.

1. INTRODUCTION

Information retrieval is the science of retrieving information relevant to information seekers from the collection of information resources such as text, images, documents, audio, as well as music. With the help of Language Technology these resources can be structured, indexed and navigated [1]. The increasing necessity of multilingual documents Retrieval in response to the user query opens up a new branch of Information Retrieval called as Cross-Language Information Retrieval. Its goal is to accept the query from user and transform that query into acceptable format which provides an interface that allows the user to retrieve and searches information in different language as per their information needs [1]. One of the key problems in IR is related to the multiple representation of a meaning. A document is retrieved for query may be different even though the term which is occurred in query and document may same. This makes difficult to match the result of relevant document against a query. This representation problem is even more evident in cross-language information retrieval (CLIR) or multi-lingual information retrieval (MLIR), where queries and documents are described in different languages. Information Retrieval engines need to convert the term in various languages in which languages users enter a query [2].

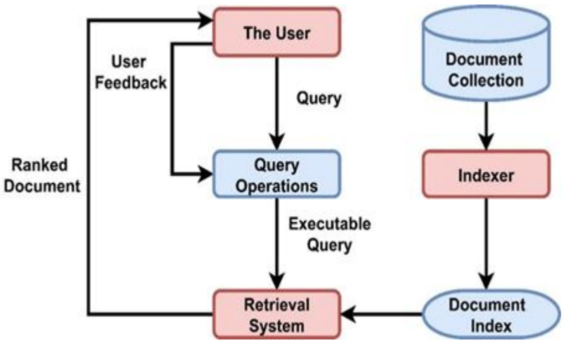

Figure 1: A General IR System Architecture [6]

Figure 1 shows the general architecture of Information Retrieval System [3], where user fires a query through operational module to IR system. Retrieval system returns documents by using indexing module which contains some query terms. The purpose of indexing module is to present the ranked documents in front of user.

The information retrieval may be classified as: Bi Lingual, Cross Lingual and Multi Lingual. Now a day, citizens of the country have more interest on global education, business and research etc. which forces them to retrieve the content from the Internet in English. But many people accessing information in a native language as they are comfortable to access the same. Cross-language information retrieval (CLIR) has a problem that is faced by certain languages that is the lack of knowledge resources of those particular languages where knowledge resources are limited such as Hindi, Malayalam and Marathi. India is a multilingual country. Indian constitution had lists of 22 languages; we referred these languages as scheduled languages. These languages are given status, recognition and official encouragement of the entire population, barely 10% Indians use English to transact and most prefer regional languages, which have evolved over centuries. As there is diversity in languages, language processing applications are a boon to the people for their day-to-day transactions [3]. India consists of the multiple states also, from all these states within the Maharashtra region specifically Marathi is used as the regional language. But Maharashtra state again divided into multiple regions such as Vidarbha, Marathwada, South Maharashtra and Konkan. The Marathi language spoken in this entire region also differs. So, finding the sense of the Marathi word is quite complicated. Marathi language consists of multiple words that are spelled same but meaning-wise/ sense-wise they are different. Such type of words when need to be from translated from source language to target Creates the problem of Ambiguity. 1. माझी पाठ दुखत आहे (My Back is aching) 2. त्याला धडा पाठ आहे (He has learnt Lesson by heart) due to this problem there is a need to use Word sense disambiguation technique to find the exact sense of the word [4]. The major contributions of this paper are summarized as follows:

- The paper presents a detailed analysis of lexical ambiguity, morphological complexity, and resource scarcity in Marathi language, highlighting their adverse impact on Marathi–English Cross-Language Information Retrieval.
- A linguistically informed hybrid CLIR model is proposed that integrates Marathi-specific morphological analysis, knowledge-based sense inventory from Marathi WordNet, contextual semantic similarity modeling, and sense-aware query reformulation.
- The study introduces a hybrid WSD mechanism that combines graph-based semantic relations with contextual similarity scoring to accurately resolve polysemy and homonymy in Marathi queries.
- Unlike conventional word-level translation approaches, the proposed method performs sense-aligned translation and weighted query expansion, significantly reducing semantic drift during cross-language retrieval.

The remainder of this paper is organized as follows. Section II reviews existing work related to cross-language information retrieval and word sense disambiguation. Section III discusses the methodological foundations of WSD approaches relevant to CLIR. Section IV presents the proposed hybrid WSD-based CLIR framework for Marathi language. Section V describes the experimental setup, datasets, and evaluation protocol. Section VI analyzes the experimental results and comparative performance. Section VII outlines the research challenges and issues in CLIR for Indian languages. Finally, Section VIII concludes the paper and discusses future research directions.

2. RELATED WORK

In this section, a work done by the research community in this area in context to word sense disambiguation is presented.

A. Query Refinement into Information Retrieval Systems

Many strategies, such search refining and significance suggestions, were implemented in an attempt to improve the effectiveness and precision of IR frameworks, particularly searches. In order to assist the framework fully comprehend the user's data wants and needs, both mechanisms—which handle the platform's imports and outputs, respectively—enter externally to the system. Query refining is a crucial strategy that can assist close this distance, particularly if the consumer is dissatisfied with the outcomes that are presented. Three factors contribute to this fulfillment: (i) the outcomes fall outside the consumer's requirements; (ii) these are too many outcomes with broad applications that make them challenging to examine; and (iii) there lack sufficient outcomes to warrant enrichment and reinforcement. Figure 2 illustrates how the search improvement, which is the focus of this study, is carried out in three distinct methods depending on the individual's degree of contribution: The cooperative approach, in which the individual chooses the words while utilizing the equipment's help by demonstrating some potential terms to add, the manual approach, in which the user fully intervenes and relies solely on his or her knowledge and personal strategies without any assistance from the tool, and the automated option, in which the machine does all the work.

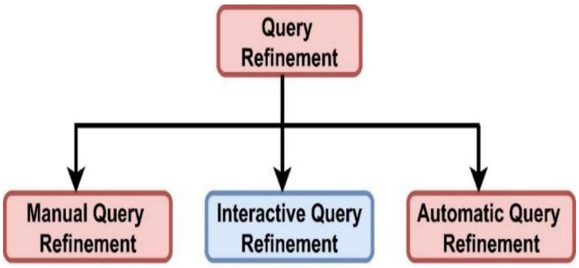

Figure 2: Categories of Query Improvement Based on the User's preferences

Paheli Bhattacharya et al. [5] proposed a clustering method based on the multilingual word vectors to group similar words across languages. For this they construct a graph with words from multiple languages as nodes and with edges connecting words with similar vectors. They use the Louvain method for community detection to find communities in this graph. They show that choosing target language words as query translations from the clusters or communities containing the query terms helps in improving CLIR. They also find that better-quality query translations are obtained when words from more languages are used to do the clustering even when the additional languages are neither the source nor the target languages.

Rabih Zbib et al. [6] proposed a neural network model to estimate word translation probabilities for Cross-Lingual Information Retrieval (CLIR). The model estimates better probabilities for word translations than automatic word alignments alone, and generalizes to unseen source target word pairs. They further improve the lexical neural translation model (and subsequently CLIR), by incorporating source word context, and by encoding the character sequences of input source words to generate translations of out-of-vocabulary words.

Vijay Kumar Sharma et al. [7] implement a dictionary-based query translation system. Queries are tokenized and multi-words query terms are created using N-gram technique. Out Of Vocabulary (OOV) terms are transliterated using the proposed OOVTTM technique. Target documents are retrieved using vector space retrieval model. Experiment results shows that the proposed approach achieves better results.

D Thenmozhi [8] proposed a methodology for Tamil–English CLIR system by translating the Tamil query to English and retrieve pages in English to address these issues. This approach uses a word sense disambiguation module to resolve the ambiguity in Tamil query. An automatically constructed ontology in English is used to address the ambiguity of English query. Authors have developed a morphological analyzer for Tamil language, Tamil–English bilingual dictionary and named entity database to translate a Tamil query to English. The translated query is reformulated using ontology and the reformulated queries are given to a search engine to retrieve English documents from the Internet.

Nazreena Rahman et al. [9] a query-based text summarization method is proposed based on common sense knowledge and word sense disambiguation. Common sense knowledge is integrated by expanding the query terms. It helps in extracting main sentences from text document according to the query. Query-based text summarization finds semantic relatedness score between query and input text document for extracting sentences.

Bilel Elayeb [10] overviewed supervised, unsupervised, semi-supervised and knowledge-based approaches. The evaluation of Arabic WSD systems is discussed in this respect. Arabic morphological analysis still suffers from the difficulty of the disambiguation challenge.

Sreelakshmi Gopal et al. [11] implemented a Supervised Malayalam word sense disambiguation system using Naive Bayes classifier. Word Sense Disambiguation is a complex problem in NLP because a particular word may have different meanings in different situations. For all human beings it is very easy to find out the accurate sense in a particular context but for machines it is very difficult to predict. Some extent of intelligence may add to the machine for an accurate prediction.

Krishnanjan Bhattacharjee et al. [12] put forward gap analysis in surveyed WSD systems comparing strengths and weaknesses of various surveyed systems and their accuracy. Based on the findings, a future hybrid approach synergizing rule-based and machine learning based methods are template. All major natural languages of the world have an intrinsic semantic feature called polysemy; same word has multiple meanings as per contexts. Word Sense Disambiguation is required by human cognition as part of Natural Language Understanding (NLU) to determine an appropriate meaning of ambiguous words in a specific context.

Alok Ranjan Pal et al. [13] proposed a model that disambiguates the actual sense of an ambiguous word in a particular context using Naïve Byes probability distribution. Authors have implemented their work using the Naïve Bayes probabilistic model.

Varinder Pal Singh et al. [14] proposed two deep learning techniques multilayer perceptron and long short-term memory (LSTM) which has been individually inspected on the word vectors of 66 ambiguous Punjabi nouns for an explicit WSD system of Punjabi language. The inputs to the deep learning techniques are the simple word vectors derived directly from manually sense-tagged corpus of Punjabi language. The multilayer perceptron has outperformed the LSTM deep learning technique for WSD task of Punjabi language. Six traditional supervised machine learning techniques have also been tested on same dataset using unigram and bigram feature sets.

Alok Ranjan Pal et al. [15] proposed an approach which uses the supervised methods as the baseline strategy for sense classification. These algorithms are tested on 13 mostly used ambiguous words. The data sets are prepared from the Bengali corpus and the Bengali WordNet. An attempt is made in this paper to report how a supervised methodology has been adopted for the task of Word Sense Disambiguation (WSD) in Bengali with necessary modifications. At the initial stage, four commonly used supervised methods, Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN) and Naive Bayes (NB), are developed at the baseline. These algorithms are applied individually on a data set of 13 most frequently used Bengali ambiguous words. On experimental basis, the baseline strategy is modified with two extensions: (a) inclusion of lemmatization process into the system and (b) bootstrapping of the operational process. As a result, the levels of accuracy of the baseline methods are slightly improved, which is a positive signal for the whole process of disambiguation as it opens scope for further modification of the existing method for better result. Authors state that the data sets are prepared from the Bengali corpus, developed in the Technology Development for Indian Languages (TDIL) project of the Government of India and from the Bengali WordNet, which is developed at the Indian Statistical Institute, Kolkata. The paper reports the challenges and pitfalls of the work that have been closely observed during the experiment.

Lokesh Nandanwar [16] proposed the graph-based unsupervised Word Sense Disambiguation Algorithm to resolve the ambiguity of a word in a given HINDI Language sentence. Finding the proper meaning of a word here implies identification of the most important node from the set of graph nodes which are representing the senses. They make use of HINDI WordNet developed at IIT Bombay as reference library of words to form the sense graph. Graph based approaches in Natural Language Processing (NLP) involves the selection of best suitable candidates (node) from many interrelated candidates. In this method, graphs correspond to senses and edges corresponds to sense relations. Authors have used the HINDI WordNet to carry out the word sense disambiguation task in NLP.

Gauri Dhopavkar et al. [17] described a rule-based method used for performing Word Sense Disambiguation task of Text in Marathi Language. In Marathi language which is spoken in Maharashtra state of India, many words are spelled same but semantically (meaning-wise/ sense-wise) different. Such words when need to be from translated from source language to target lead to ambiguity. This paper states that their method successfully identifies the correct sense of the given text from the predefined possible senses using word rules and sentence rules. The system presented works on only single sentence and identifies the ambiguity. The system accuracy is around 75% which include disambiguation of nouns, adjectives and verbs in Marathi language. Authors have stated that the system can only identify and resolve word level ambiguity.

Nutan B. Zungre et al. [18] proposed a Graph-based algorithm, through which word ambiguity is resolved based on their senses and context domain. In Graph-based algorithm, a graph is created that comprises the word which is to be disambiguated with their corresponding candidate sense. In the proposed work, sense disambiguation has been done for Marathi language words. Through Marathi WordNet prepared by IIT-Bombay, multiple senses of Marathi word have been explored. The proposed system is used to figure out the rightful sense of word in Marathi language using Decision graph algorithm. The proposed system uses Source Language as Marathi. Input Text in Marathi is obtained through Google Input Tool. Marathi WordNet is used to obtain the exact features of each word in the sentence. Marathi language is the Target language for which the sense disambiguation has been executed.

Sudha Bhingardive et al. [19] presented the usage of various features of Indo WordNet in performing WSD. Indo WordNet is a linked structure of Wordnets of major Indian languages. Several Indo WordNet-based WSD approaches have been proposed and implemented for Indian languages. Author stated that the unsupervised approaches are better alternatives than supervised approaches as they do not require any sense-annotated corpora whose creation needs lots of manual efforts.

R. K. Sharma et al. [20] describes the process of creation of Punjabi WordNet, where semantic relations are borrowed from the Hindi language, while the lexical relations are created for Punjabi language, as these relations are language dependent. To create lexical relations, a lexical relation tool has been proposed in this paper. The development of bilingual Hindi–Punjabi and Punjabi–Hindi dictionaries through this process has also been presented in this paper. This paper also discusses the challenges in the development of Punjabi WordNet and creation of language-specific synsets with reference to Punjabi WordNet. The expansion approach for the development of Indo WordNet has made this development process fast. The semantic relations are borrowed from the source language, Hindi, as they are same for all the languages. Lexical relations are language specific, so they cannot be borrowed from the source language. Authors have put forward their observation that by using lexical creation tool, the creation of lexical relations has become very fast and easy particularly for Hindi in-family languages, while for languages that do not fall in the same family, the provision of creation of lexical relation without referring to Hindi WordNet has been very useful. The Indo WordNet developed through this initiative will serve as an important tool in the field of Natural language processing for Indian languages.

Neeraja Koppula et al [21], proposed a WSD system for regional Telugu language. Word sense disambiguation system can be developed using three approaches; in this work, we are using knowledge-based approach, where the accuracy is more than unsupervised approaches. In Regional Telugu language, proposed works are WFS (Word First Sense) finding the correct sense of the polysemy word, by assigning the first sense of the polysemy word from the LKB (Lexical Knowledge Base) as an appropriate sense and WMFS (Word Most Frequent Sense) is the process of identifying the sense, having a greater number of context words that sense is treated as Most Frequent Sense of the polysemy word. These two methods are context independent, and the proposed method WTSS is context dependent. The WFS and WMFS methods are related to the proposed method Word Total Sense Score (WTSS) in this paper. The proposed algorithm word total sense score (WTSS) is using knowledge-based approach for word sense disambiguation in regional Telugu language.

Sanjay Kumar Dwivedi et al [22] discussed question classification word sense disambiguation. The automated translation of question papers from English to Hindi is one such key area which requires suitable WSD techniques to resolve ambiguity in a question word. When machine translates question sentences, it faces ambiguity problem that results in ambiguous translation. Identification of question type is important for remove ambiguity in the question paper. This paper reviews many types of question sentences and tries to identify their pattern and discusses how to disambiguate their meaning.

3. METHODOLOGY

In this section, based on the literature surveyed, following methodologies have been identified with respect to Word Sense Disambiguation.

A. Dictionary-based

The most important aspect to cross-lingual IR is to replace each query term with most appropriate translations extracted automatically from Machine Readable Dictionaries (MRD). The translation which was made using bilingual dictionaries is seems to be simple but Ballesteros and Croft [23] perform the experiment and concluded 40-60% loss in effectiveness when compared to monolingual retrieval.

B. Parallel corpus

A parallel corpus is a collection where texts in one language are aligned with their translations in another language. Several systems have been developed to mine large parallel corpora from the web. Wang and Lin [2010] gave a method which first identifies a set of seed URLs and crawl candidate bilingual websites. The obtained pages are cleaned and bilingual texts collected to construct comparable corpora. The bilingual search result pages obtained from a real search engine as a corpus for automatic translation of unknown query terms not included in the dictionary [24]. They propose a PAT-tree based local maxima method for effective extraction of translation candidates. The approach gives excellent results.

C. Comparable corpus

Comparable corpus, on the other hand, consist of texts that are not translations, but share similar topics. for example, newspaper collections written in the same time period in different countries. Sadat Fatiha [25] exploited the idea of using multilingual based encyclopedias such as Wikipedia to extract terms and their translations to construct a bilingual ontology or enhance the coverage of existing ontologies. The method shows promising results for any pair of languages. Qian & Meng [26] expanded Chinese OOV phrase with its partial English translation and submitted to the search engine. The translation of OOV words is mined by preprocessing the snippets obtained to extract the main text from the web page. The strings obtained are sorted by weighted frequency to output the top n translation of OOV phrase. The method proves to obtain the translation with high time efficiency and high precision.

D. LSA-WSD (Latent Semantic Analysis-Word Sense Disambiguation)

The LSA-WSD system uses a modified clustering technique (SMC) that exploits the properties of the LSA cognitive model for relating semantically similar items to perform sense discovery. It is based on the concept that each term vector in an LSA semantic space carries all the possible meanings or senses for a term. The challenge for sense discovery is then to separate these senses into individually identified senses. rs.

E. Proposed WSD-Based CLIR model for Marathi Language

Marathi, as a morphologically rich and polysemous Indian language, poses significant challenges for Cross-Language Information Retrieval (CLIR), particularly due to lexical ambiguity, dialectal variations, and limited linguistic resources. Conventional dictionary-based query translation often fails to preserve semantic intent, leading to degraded retrieval precision. To address these limitations, this study proposes a hybrid Word Sense Disambiguation–driven CLIR model specifically designed for Marathi–English information retrieval. The framework integrates linguistic preprocessing, knowledge-based disambiguation, contextual semantic modeling, and weighted query reformulation to ensure semantically faithful translation and improved retrieval effectiveness.

Dataset and Corpus Description

To evaluate the effectiveness of the proposed WSD-based CLIR model, experiments were conducted on a Marathi–English cross-language retrieval task. The experimental setup consists of a Marathi query set and an English document corpus. Marathi queries were designed to include a high proportion of ambiguous and polysemous words commonly observed in real-world information needs. The English document collection covers multiple domains, including news articles, educational content, and general web documents, ensuring diversity in vocabulary and semantics.

Marathi lexical resources, including Marathi WordNet, were employed for sense inventory generation and semantic relation extraction. For translation and evaluation purposes, a bilingual Marathi–English lexicon was utilized. All documents were preprocessed using standard information retrieval techniques, including tokenization, stop-word removal, and TF–IDF-based indexing.

Marathi Query Processing: User queries are accepted in Marathi script and undergo initial normalization, including Unicode normalization and punctuation removal. Tokenization is performed using rule-based segmentation to preserve compound words and postpositions, which are common in Marathi syntax. Stop-word removal is applied selectively to avoid eliminating semantically significant functional markers.

Morphological and Linguistic Analysis: Given the highly inflectional nature of Marathi, a morphological analyzer is employed to extract root forms, grammatical categories, and inflectional attributes (tense, gender, number, case). This step reduces lexical sparsity and enables accurate sense mapping by collapsing multiple surface forms into a single canonical lemma.
Let a query $Q = \{w_1, w_2, \dots, w_n\}$
Each word wi is transformed into its lemma li using:

$$l_i = Morph(wi)$$

where $Morph(\cdot)$ denotes the morphological normalization function.

Candidate Sense Generation: For each lemma li, a set of possible senses is extracted from Marathi WordNet, including glosses, example sentences, semantic relations (synonymy, hypernymy), and domain labels. This forms the candidate sense set:

$$S_i = \{s_{i1}, s_{i2}, \dots, s_{ik}\}$$

where each sij represents a potential sense of word l_i.

Hybrid Word Sense Disambiguation: To robustly resolve ambiguity, a hybrid WSD strategy is adopted, combining knowledge-based graph modeling with contextual semantic similarity.

4.1 Sense Graph Construction: A sense graph $G = (V, E)$ is constructed where, Nodes V represent candidate senses. Edges E represent semantic relations derived from WordNet (synonymy, hypernymy, co-occurrence)

Contextual Similarity Scoring: The local query context is represented using contextual embeddings. Each sense gloss is embedded in the same semantic space. The similarity score between sense sij and query context C is computed as:

$$Sim(sij, C) = cos(sij, C)$$

Sense Selection: The final sense is selected by maximizing a joint score combining graph centrality and contextual similarity:

$$s_i^* = args \max_{s_{ij} \in s_i} (\alpha . Centrality(s_{ij}) + \beta . sim(s_{ij}, C))$$

where α and β are tunable weights satisfying $\alpha + \beta = 1$

Sense-Aware Query Translation: Once the optimal sense s_i^* is identified, translation is performed using sense-aligned bilingual mappings rather than direct word translation. This ensures that only semantically valid English equivalents are selected, reducing noise caused by polysemy and homonymy.

Each translated term is assigned a confidence weight proportional to its disambiguation score:

$$w_i = Score(s_i^*)$$

Semantic Query Expansion and Reformulation: To improve recall, the translated query is expanded using:

- Synonyms and hypernyms of the selected sense
- Domain-specific related terms

The final reformulated query Q' is expressed as a weighted vector:

$$Q' = \{(t_1, w_1), (t_2, w_2), \ldots, (t_m, w_m)\}$$

where t_i are English terms and wi are their normalized semantic weights.

Cross-Language Document Retrieval and Ranking: The reformulated query is submitted to an English document collection indexed using a vector space or probabilistic retrieval model. Documents are ranked based on cosine similarity or relevance probability, incorporating query term weights to prioritize semantically consistent matches.

4. RESULT ANALYSIS

This section presents the performance of the proposed model. The experimental evaluation of the proposed model carried out in google colab environment using Python programming.

TABLE 1: COMPARATIVE PERFORMANCE ANALYSIS OF PROPOSED MODEL WITH EXISTING MODELS

Model	**P@10**	**P@20**	**Recall**	F1	MAP	nDCG@10
Dictionary based CLIR	0.42	0.39	0.58	0.47	0.36	0.44
First Sense Baseline	0.48	0.45	0.62	0.53	0.41	0.50
Knowledge based WSD	0.52	0.5	0.66	0.58	0.46	0.56
Proposed Hybrid WSD-based CLIR	0.62	0.57	0.71	0.65	0.54	0.64

Table 1 shows the comparative analysis demonstrate the progressive improvement in retrieval effectiveness as semantic awareness is incorporated into the CLIR pipeline. The dictionary-based CLIR approach yields the lowest performance across all metrics (P@10 = 0.42, MAP = 0.36), reflecting significant semantic drift caused by unresolved lexical ambiguity. Introducing the First Sense Baseline leads to moderate gains in precision and ranking quality, indicating that sense frequency partially mitigates ambiguity but remains context-insensitive. The knowledge-based WSD model further improves performance (P@10 = 0.52, MAP = 0.46, nDCG@10 = 0.56) by leveraging semantic relations from lexical resources, thereby enhancing contextual relevance. The proposed Hybrid WSD-based CLIR framework achieves the best results across all evaluation metrics, with notable improvements in early precision (P@10 = 0.62), recall (0.71), and ranking effectiveness (MAP = 0.54, nDCG@10 = 0.64). These results confirm that integrating hybrid word sense disambiguation with sense-aware query reformulation significantly enhances both the accuracy and ranking quality of Marathi–English cross-language information retrieval.

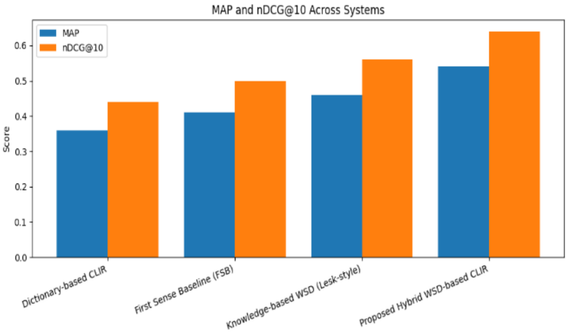

Figure 3: Comparative analysis of MAP and nDCG@10 for different CLIR model

Figure 3 presents a comparative analysis of retrieval effectiveness in terms of Mean Average Precision (MAP) and nDCG@10 across different CLIR approaches. The dictionary-based CLIR system exhibits the lowest MAP and nDCG@10 scores, indicating weak ranking quality due to unresolved semantic ambiguity during query translation. The First Sense Baseline shows a modest improvement, suggesting that selecting the most frequent sense partially enhances relevance but remains insensitive to contextual cues. Incorporating knowledge-based WSD further improves both MAP and nDCG@10, reflecting more accurate semantic alignment between queries and retrieved documents. The proposed Hybrid WSD-based CLIR framework achieves the highest scores for both metrics, demonstrating superior ranking performance and early precision. This improvement confirms that integrating contextual and semantic disambiguation mechanisms significantly enhances the quality and ordering of retrieved documents in cross-language information retrieval.

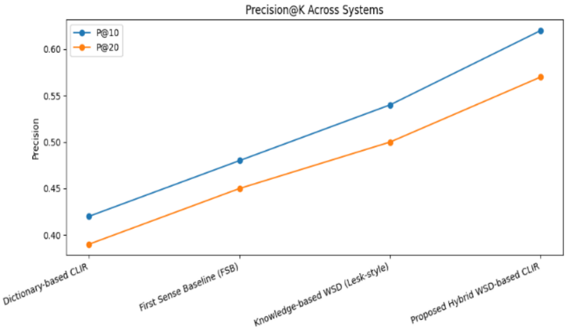

Figure 4: Comparison of Precision@10 and Precision@20 across different CLIR model

Figure 4 shows the variation of early precision metrics, Precision@10 (P@10) and Precision@20 (P@20), across different CLIR systems. The dictionary-based CLIR approach yields the lowest precision values, indicating limited effectiveness in retrieving relevant documents at top ranks due to ambiguous word translations. The First Sense

Baseline demonstrates a noticeable improvement in both P@10 and P@20, reflecting partial mitigation of ambiguity through frequent-sense selection. The knowledge-based WSD method further enhances precision by exploiting semantic relations, leading to more contextually relevant retrieval. The proposed Hybrid WSD-based CLIR framework achieves the highest precision at both cutoff levels, with a pronounced gain in P@10, highlighting its superior capability to rank highly relevant documents within the top results. These trends confirm that sense-aware and context-driven disambiguation substantially improves early retrieval precision in cross-language information retrieval.

5. Research issues

Following are the issues observed with respect to cross language information retrieval while going through literature survey:

Homonymy: The words having two or more different meaning.

Polysemy: The word having multiple related meaning.

Word inflection: They may have different grammatical forms.

Phrase translation: Phrase gives different meaning than the words of phrase.

Lack of resources: Unavailability of regional language resources for experimentation.

6. Current Challenges in IR

In CLIR, there is much more serious problem than that of monolingual retrieval where one language can be used for query and document retrieval. When compared with the monolingual retrieval may have some of the challenges in cross languages information retrieval. Finding or producing information and assets for the Hindi and Marathi languages is difficult. These days, across these resource-constrained languages, there exists just a few pairings of searches and pertinent resources for cross-language retrieving. But still a hundred times more specific data for particular area is needed and it is crucial to find an effective and inexpensive way to collect these additional data and appropriate resources to be used for evaluations of cross-lingual IR for these languages. [28]

A. Language Resources:

CLIR requires support for language resources such as bilingual dictionaries, parallel corpora and comparable corpora. At present mainstream language only have popular resources and very few resources for low resources languages. Therefore, it is required to build high-quality language resources.

B. Ambiguity

Uncertainty may arise whenever a phrase is overlapping during that moment. For instance, the term "maan," which means "respect" or "neck," contains two different interpretations in the same language Marathi as neck and respect. When translating such a word into other languages, because of the polysemy, it is needed to combine more contexts to improve accuracy [29].

C. Phrase identification and translation

Identifying phrases in limited context and translating them as a whole entity rather than individual word translation is difficult [28].

D. Translation

Certain words are unclear and require transliteration rather than interpretation. For instance, in Marathi, □□□□□□ (Bhaskar, Sun) denotes both the sun and the name of an individual. Detecting these cases based on available context is a challenge [30].

E. Transliteration errors

Errors while transliteration might end up fetching the wrong word in target language [31].

F. User Feedback

Users have careful about their information needs. If user must be satisfied with the documents retrieved [32].

Conclusion

This study addressed the critical challenge of lexical ambiguity in Marathi–English Cross-Language Information Retrieval by systematically integrating Word Sense Disambiguation into the query translation and retrieval pipeline. Through an extensive review of existing CLIR and WSD approaches, the limitations of traditional dictionary-based and frequency-based translation methods were identified, particularly in handling polysemy, homonymy, and morphological variation in Marathi language. To overcome these challenges, a hybrid WSD-based CLIR framework was proposed, combining morphological normalization, knowledge-based semantic modeling using Marathi WordNet, contextual similarity analysis, and sense-aware query reformulation. Experimental evaluation using standard retrieval metrics demonstrated consistent and significant improvements over baseline approaches, especially in early precision and ranking quality, as reflected by higher P@10, MAP, and nDCG@10 scores. The results confirm that resolving word sense ambiguity at the query level plays a pivotal role in improving semantic alignment between Marathi queries and English documents, thereby enhancing overall retrieval effectiveness. The proposed framework establishes a robust and linguistically informed foundation for CLIR in low-resource Indian languages.

Several directions remain open for future research. First, the proposed framework can be extended by incorporating deep contextualized multilingual embeddings and transformer-based models to further improve sense representation and disambiguation accuracy. Second, phrase-level and multi-word expression disambiguation can be explored to address limitations arising from compound terms and idiomatic expressions in Marathi. Third, the framework can be generalized to support multiple Indian languages by using Indo

WordNet and shared semantic representations. Additionally, integrating user relevance feedback and adaptive learning mechanisms may further enhance retrieval performance over time. Finally, large-scale evaluation on diverse real-world datasets and domain-specific corpora would strengthen the applicability and scalability of the proposed approach for practical deployment in multilingual information retrieval systems.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

FUNDING SUPPORT

The author declare that they have no funding support for this study.

REFERENCES

[1] HL Shashirekha and Ibrahim Gashaw, "Enhanced Amharic-Arabic Cross-Language Information Retrieval System using Part of Speech Tagging", IEEE, 2019.

[2] Nurul Amelina Nasharuddin et al, "A Review on Building Bilingual Comparable Corpora for Resource-limited Languages", IEEE, Fourth International Conference on Information Retrieval and Knowledge Management, 2018.

[3] Jay Patel et al, "Cross-lingual Information Retrieval: application and Challenges for Indian Languages", 5th International Conference for Convergence in Technology (I2CT) Pune, India, Mar 2019.

[4] Gauri Dhopavkar et al, "Application of Rule Based Approach to Word Sense Disambiguation of Marathi Language Text", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and communication systems (ICIIECS) 978-1-4799-6818-3/15, 2015.

[5] Paheli Bhattacharya et al, "Using Communities of word Derived from Multilingual Word Vectors for Cross-Language Information Retrieval in Indian Languages", ACM Trans. Asian Low-Resour. Lang. Inf. Process. 18, 1, Article 1, December 2018.

[6] Rabih Zbib et al, "Neural-Network Lexical Translation for Cross-lingual IR from Text and Speech", ACM, ISBN 978-1-4503-6172-9, 2019.

[7] Vijay Kumar Sharma, "Cross-Lingual Information Retrieval: A Dictionary-Based Query Translation Approach", Springer, Advances in Computer and Computational Sciences, Advances in Intelligent Systems and Computing 554, 2018.

[8] D Thenmozhi, "Ontology-based Tamil–English cross-lingual information retrieval system", Springer, June 2018.

[9] Nazreena Raman and Bhogeswar Borah, "Improvement of query-based text summarization using word sense disambiguation", Complex & Intelligent Systems (2020) 6:75–85, Springer, 2020.

[10] Bilel Elayeb, "Arabic word sense disambiguation: a review", Springer, 12 March 2018.

[11] Sreelakshmi Gopal and Rosna P Haroon, "Malayalam Word Sense Disambiguation using Naïve Bayes Classifier", IEEE, International Conference on Advances in Human Machine Interaction (HMI - 2016), March 2016.

[12] Krishnanjan Bhattacharjee et al, "Survey and Gap Analysis of Word Sense Disambiguation approaches on Unstructured Texts", Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE, ISBN: 978-1-7281-4108-4, 2020.

[13] Alok Ranjan Pal and Diganta Saha, "Word Sense Disambiguation in Bengali: An Auto-updated Learning Set Increases the Accuracy of the Result", Springer Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing, pp. 423-430, 2016.

[14] VARINDER PAL SINGH, "Word sense disambiguation for Punjabi language using deep learning techniques", Springer, November 2019.

[15] ALOK RANJAN PAL, "Word Sense Disambiguation in Bangla Language Using Supervised Methodology with Necessary Modifications", Springer, 11 May 2018.

[16] Lokesh Nandanwar, "Graph Connectivity for Unsupervised Word Sense Disambiguation for HINDI Language", IEEE Sponsored 2nd International Conference on Innovations in Information Embedded and Communication systems (ICIIECS), 978-1-4799-6818-3, 2015.

[17] Gauri Dhopavkar, "Syntactic Analyzer using Morphological Process for a Given Text in Natural Language for Sense Disambiguation", IEEE 978-1-4799-4236-7, 2014.

[18] Nutan B. Zungre, Gauri M. Dhopavkar, "Sense Disambiguation For Marathi Language Words Using Decision Graph Method", 978-1-4673-9214-3/16 , IEEE Sponsored World Conference on Futuristic Trends in Research and Innovation for Social Welfare(WCFTR),2016.

[19] Sudha Bhingardive and Pushpak Bhattacharyya, "Word Sense Disambiguation Using Indo WordNet", Springer, pp. 243-260, 2017.

[20] R.K. Sharma and Parteek Kumar, "Development of Punjabi WordNet, Bilingual Dictionaries, Lexical Relations Creation and Its Challenges", Springer, The WordNet in Indian Languages, pp. 83-99, 2017.

[21] Neeraja Koppula et al, "Word Sense Disambiguation in Telugu Language Using Knowledge-Based Approach", Springer, Proceedings of the Third International Conference on Computational Intelligence and Informatics, Advances in Intelligent Systems and Computing, Vol. 1090, pp. 153-161, 2020.

[22] Sanjay Kumar Dwivedi and Shweta Vikram, "Word Sense Ambiguity in Question Sentence Translation: A Review", Springer, Information and Communication Technology for Intelligent Systems (ICTIS 2017), Volume 84, pg. no. 64-71, 2017.

[23] Ballesteros, and Croft, " Dictionary Methods for Cross-Lingual Information Retrieval", 7th DEXA Conf. on Database and Expert Systems Applications, Pg no. 791-801, 1996.

[24] Jenq-Haur Wang, "Translating Unknown Cross-Lingual Queries in Digital Libraries Using a Web- based Approach", Conference on Digital Libraries (JCDL'04), ACM, 2004

[25] Sadat Fatiha , "Exploiting a Multilingual Web-based Encyclopedia for Bilingual Terminology Extraction", PACLIC 24 Proceedings, 2011.

[26] Pratibha Bajpai, "Cross Language Information Retrieval: In Indian Language Perspective", IJRET: International Journal of Research in Engineering and Technology, June-2014