International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

A New Approach to Enhance Query Refinement for Marathi Language Information Retrieval

Vivek Ajabrao Manwar[1], Dr. A. B. Manwar[2]

[1] Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India
[2]Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India
E-mail address: [1]vivek.manwar007@gmail.com, [2] avinashmanwar@sgbau.ac.in

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: The Information Retrieval (IR) enable users to access relevant data in a language different from their queries. This involves the capability to submit a query in one language and retrieve documents in Same or another, like Marathi or English. This is accomplished by creating a system that allows for the comparison of a query in one language with data in same or another language. This research presents a novel approach to enhance query refinement for Marathi language information retrieval, focusing on addressing the language's unique linguistic complexities. The study begins with the collection and preprocessing of a Marathi dataset, emphasizing relevant features. Through comprehensive query analysis, common linguistic variations and ambiguous words in queries are identified. Using natural language processing (NLP) and semantic understanding techniques tailored for Marathi ambiguous words, a robust framework for contextual query interpretation is developed. The methodology includes preprocessing, feature extraction, developing a query refinement model, and building a machine learning model. Performance evaluation of the proposed model is conducted using various metrics such as accuracy of 0.79, mean squared error (MSE) is 0.2077, specificity of 1.0 and sensitivity of 47.54% on the refined queries. The proposed model aims to make a valuable contribution to the field of Marathi language information retrieval.

Keywords: Marathi Language Information Retrieval, Query Refinement, Natural Language Processing (NLP), Linguistic Complexity, Machine Learning Model

1. INTRODUCTION

The increase in diversity among online search users worldwide has created new challenges for search engine providers, especially when dealing with different user languages. To address this, a process called "localization" is used, where systems are adapted for multiple domains through translation and cultural adjustments. Instead of creating separate models for each market and target language, a more effective approach considers that many people globally are proficient in multiple languages, learning more than two on average [1]. This helps bridge the lexical gap between languages, a process referred to as cross-lingual information retrieval (CLIR) [2]-[3]. In traditional monolingual information retrieval (IR), connecting this lexical gap is difficult, leading to inefficient semantic normalization. A better solution involves developing semantic based implications within text embedding, allowing for the generalization of vocabularies from labeled data and improving retrieval results to overcome data sparsity issues. This model is applied in multilingual IR strategies.

To bridge the gap between languages for cross-lingual access, models using cross-lingual concepts are employed. The Full-blown Machine Translation (MT) [4] scheme is one such model used to convert queries and documents, but it requires a large amount of parallel data, which is often lacking for many languages and applications. Another model tackles the lexical gap by utilizing queries and documents with external multilingual knowledge sources like Wikipedia and Babel Net [5]. However, this approach is limited in concept coverage for resource-based languages, and the knowledge base has restricted content, a challenge addressed by the CLIR approach. Bilingual text embedding is a more versatile option compared to the two paradigms, but it faces constraints, such as the need for bilingual supervision signals to create distributed cross-lingual semantic spaces [6]-[7]. This supervision comes in the form of hybrid sentence-based

E-mail:author's email

parallel data [8], pre-defined word conversion, and user document-centric query refinement comparable data [9].
Numerous features of retrieving data make it a good fit for artificial intelligence. Relevant data retrieving is classifying activities, which up until recent needed previously been completed manually. These tasks have become appropriate to various ML models. The IR techniques can provide a wealth of instances, properties, and parameters that are used to train algorithms. The majority of these systems contain hundreds of thousands of texts, and a multitude of variables can be obtained from the many linguistic aspects which are accessible from texts. As a result, it has a lot of data accessible; become clear because typical algorithms have too many characteristics and samples for ML algorithms to handle effectively.
The contribution of study has given below

- Innovative approach to enhancing query refinement for Marathi language information retrieval.
- Addressing linguistic complexities specific to Marathi, such as ambiguous words and variations in query structures.
- Providing a valuable framework for improving the accuracy and efficiency of information retrieval systems.
- Utilizing NLP techniques tailored for Marathi, along with implementing the Rocchio algorithm and vector space model.
- Demonstrating a novel methodology for contextual query interpretation and refinement.
- Advancing the field of Marathi language processing.
- Setting a foundation for future research in improving information retrieval systems for other less-resourced languages.

Organization of the paper

Section 2 discussed the previous studies and their methodologies. Section 3 presented the methodology of Development of a query refinement model using the vector space model and Rocchio algorithm. Section 4 presented the result analysis of the proposed model. Finally, conclusion and future scope were discussed in section 5.

2. RELATED WORK

Systems for storing and finding information help people access large amounts of information [10]-[11]. When someone wants to learn something, they use these systems. The person (called the user) tells the system what they're looking for, like a list of keywords or maybe a sample document [12]. The system then looks in its database for documents that match the user's request and shows the ones that are most relevant. This is generally about information retrieval and how computers implement Multilingual retrieval systems.
Recent study shows significant efforts to overcome language barriers when searching for information on the Internet, irrespective of the language used to write the content. A few investigators [13]-[14]-[15] have looked into the scope of Multilingual Information Retrieval (IR) from various angles, while others [16]-[17]-[18] have suggested entirely new frameworks and models for Multilingual Information Retrieval. Many Multilingual Retrieval systems have been proposed for different language pairs, making it possible to access information on the Internet, no matter the language it was originally written in English or Marathi.
Researchers have been actively studying multilingual retrieval models in the information retrieval field. One important study, [19], suggests a novel approach to information retrieval by treating it. This addresses the statistical translation issue and suggests a methodology to improve the efficiency of retrieving. A novel approach to estimating statistical translating frameworks for customized data retrieval that utilizes mutual data appears by [20], additional noteworthy addition. [21] describes a novel method for estimating translation models during search engines by using access data. In addition to this, several works on translating a language framework for retrieving data contain theoretical considerations. For example, [22] offers an ontological examination of the framework, going over its characteristics and constraints. In order to improve retrieval efficiency, [23] offers a novel technique for search development that makes use of multilingual SMT. In the same way, [24] presents a novel method of query expanding that enhances retrieving performance by making use of searching records and conceptual translating frameworks.
Researchers have suggested using neural-based methods, like CNNs and RNNs, to make information retrieval systems work better. These techniques entail using a continuous vector space for modeling text files and search queries. It makes it possible to order texts according to how relevant one appear to a particular query using neural network-based metrics such cosine similarities [25]. In addition to learning intricate non-linear patterns for text input, such neural techniques can enhance retrieving methods [26]. Furthermore, attention-based strategies consisting of understanding the Transformer framework [27] performed to improve the ability of IR algorithms to concentrate on relevant portions of the queries and texts to compare. them. Furthermore, it has been demonstrated that utilizing pre-

trained language models, like BERT [28] and GPT-2 [29], improves the effectiveness of IR algorithms by offering researchers a greater comprehension of the contextual and significance of natural language searches and texts.

One method, called "DeepTR" in [30], learns to change the importance of words based on how similar they are in meaning using distributed representations. Another approach, explained in [31], involves adding neural word meanings into mechanisms for retrieving data. The authors suggest a way to measure how well different neural word meanings work in data retrieving and explain how to use them in retrieval models. Methods like "DeepCT" from [32], which considers the context, try to make the first-stage retrieval process better by giving more importance to words that are more relevant based on their context. The researchers propose an approach to DL which discovers important words in a sentence based on their meaning. Another approach called "TDV" in [33] concentrates on learning how words can be told apart in information retrieval.

In document search frameworks, embedded words have gained popularity as a means of representing texts and query. These persistent, intricate approximations of phrases that convey their significance are called embedded data. For instance, Word-Embedding inspired [34] represents texts and query using embedded words and the Fisher Vector technique. Bilingual embedded words are used in [35] to facilitate cross-lingual content retrieving. uses embedded words in [36] for expressing brief texts. In [37], texts and questions are represented using two distinct embedded spaces; in [38], embedded words are used to display context and possible answers for natural language production. Neural networks are used in [39] and [40] to arrange content and query in a continous space, and a similarity metric is used to determine the order of the data. In [41] improves retrieving performance by organizing key terms in texts using a neural network. Using algorithms based on transformers that have already been trained to encode texts and queries is another well-liked method. offers a technique for text retrieving in [42] that uses a transformer-based model that has been taught to encode queries and texts independently, a process referred to as "DC-BERT."

Word-based embedding [35] proposed a method for improving retrieving data by combining multilingual and cross-lingual embedded data. introduced a GLM that improves retrieving effectiveness by utilizing embedded words [43]. To enhance retrieving effectiveness, a technique for combining embedded words through set procedures was presented in [44]. For improved retrieving effectiveness, a method for merging embedded words into a multiple embedded space modeling was proposed in [37]. In order to enhance retrieving effectiveness, [45] proposed using k-NN searching in place of term-based retrieving and including translation frameworks and BM25. [46] developed a technique to improve query responding by combining CNN and BOW. presented the DenSPI technique in [47] to enhance real-time query responding effectiveness. A technique known as SPARC was introduced in [48] to enhance real-time query responding effectiveness. CoRT, which combines BM25 through transformer-based scenarios, was presented in [41] in an effort to enhance retrieving effectiveness. To improve retrieving effectiveness, [49] introduced MEHybrid, which combines attentional techniques using sparse and dense models. introduced the CLEAR in an effort to enhance retrieving effectiveness by fusing semantically and vocabulary residual embedded data. To increase retrieval of documents from system memory, a hybrid strategy combining semantics and vocabulary searching was presented in [51].

In summary, these papers demonstrate various ways to make text retrieval systems work better and faster. They suggest different approaches like adjusting relevance weights in context, predicting trade-offs between efficiency and effectiveness, creating quick access indexes, making sparse representations more detailed, and using a unified approach for large-scale retrieval.

3. PROPOSED METHODOLOGY

The proposed model aims to enhance query refinement for Marathi language information retrieval by addressing linguistic complexities unique to the language. The study is begun with collecting and preprocessing a Marathi dataset, considering relevant features. Through a comprehensive query analysis, identify common linguistic variations and ambiguous words in queries. Leveraging natural language processing (NLP) and semantic understanding techniques tailored for Marathi ambiguous word, develop a robust framework for contextual query interpretation for ambiguous word. The methodology includes preprocessing, feature extraction, develop query refinement model, build ML model. The performance of proposed model can be measure by various evaluation metrics such as Accuracy, precision, recall, and F1 score has been employed of the refined queries, with the goal of providing a valuable contribution to the field of Marathi language information retrieval. Figure 1 shows the architectural diagram of proposed query refinement model.

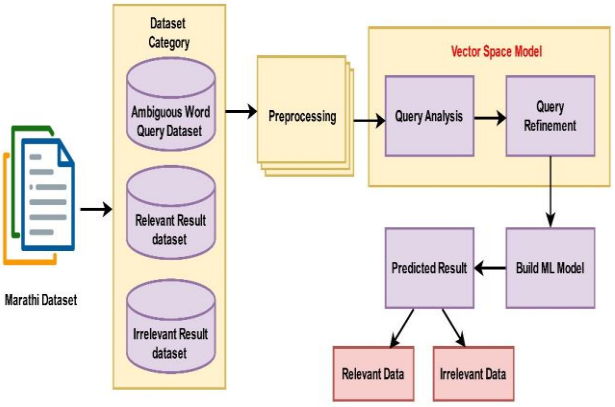

Figure 1: Proposed Architecture of Query Refinement

A. Dataset Description

The research work necessitates a standard benchmark dataset; however, as of the undertaken research work to date, no suitable dataset is available due to the ambiguous nature of queries in the Marathi language. To address this, a dataset has been utilized, collected from various news websites and official sources such as https://marathivishwakosh.org for information retrieval in Marathi. The dataset is planned to be categorized into three parts: one containing ambiguous queries and their respective domain-specific query data, another containing relevant result data, and the third containing irrelevant result data. The Query dataset comprises 215 ambiguous words, while the results parts contain approximately 4440 records as shown in the table 1, 2, 3. The model is trained using 70% of the dataset and tested with the remaining 30%.

Table 1: Ambiguous Word Query Dataset with sample queries

Sr. No.	Name of Dataset	Ambiguous Word	Query Data
1	Que_Ref_WSD_Que	भाव	1. सोयाबीनला योग्य भाव मिळत नसल्याने शेतकरी चिंतेत - दर, किंमत 2. सुखद वा असुखद भाव जाणवत असतो. - भावना, मनोविकार 3. चेहऱ्यावर आश्चर्याचा भाव - चेहऱ्यावर उमटणारा मानसिक स्थितीचा निदर्शक असा विशेष
2		भरती	1. सक्तीच्या सैन्य भरती चा प्रकार प्रशिया - सैन्य, नोकरी इत्यादीत घेण्याची किंवा प्रवेश मिळवण्याची क्रिया 2. पृथ्वीच्या बहुतेक भागांत दिवसाकाठी दोन वेळा भरती व दोन वेळा ओहोटी येते - जगातील प्रमुख ठिकाणांवरील भरती- उधाण 3. खेडयांतून शहरांत स्थलांतर करणाऱ्यांत पुरुषांचा भरणा(मोठ्या प्रमाणावर असलेला समावेश-भरणा)

Table 2: Relevant Result dataset with sample data

Sr. No.	Name of Dataset	Sample Data
1	Que_Ref_WSD_txt	1.1. सोयाबीनच्या भावात अजूनही चढ उतार सुरू आहेच. सोयाबीनला चांगला भाव मिळण्याची वाट शेतकरी पाहतो आहे. बाजारातील मालाची आवक आणि भाव यांचा थेट संबंध असतो. परिणामी एखादे पीक जोरात आले की त्याची बाजारातील आवक वाढते आणि मग त्या पीकाचा भाव कोसळतो. यावेळेस मात्र सोयाबीनला भाव मिळत नसल्याने शेतकरी बाजारात बाजारात सोयाबीन आणण्याचे टाळत आहे. तो भाव वाढण्याची वाट पाहतो आहे. 1.2. सोयाबीनचे (Soybean)पीक हाती येऊनदेखील शेतकरी हवालदिल आहे. कारण सोयाबीनच्या भावात (Soybean rate)अजूनही चढ उतार सुरू आहेच. सोयाबीनला चांगला भाव मिळण्याची वाट शेतकरी पाहतो आहे. बाजारातील मालाची आवक आणि भाव यांचा थेट संबंध असतो. परिणामी एखादे पीक जोरात आले की त्याची बाजारातील आवक वाढते आणि मग त्या पीकाचा भाव कोसळतो. यावेळेस मात्र सोयाबीनला भाव मिळत नसल्याने शेतकरी (Farmers) बाजारात बाजारात सोयाबीन आणण्याचे टाळत आहे. तो भाव वाढण्याची वाट पाहतो आहे. त्यामुळे आगामी दिवसांमध्ये सोयाबीनचे भाव कोणत्या पातळीवर स्थिरावतात याकडे सगळ्यांचे लक्ष असणार आहे. भाव (फीलिंग) आणि मनोभाव अथवा भावना या दोहोंमध्ये काही बाबतीत साम्य असले, तरी महत्वाचा भेदही दिसून येतो. सुखद व असुखद भावांप्रमाणेच प्रेम, भय, क्रोधादी मनोभाव व्यक्तींच्या अभिसरणरूप अथवा अपसरणरूप प्रतिक्रियांना कारणीभूत होत असतात. भाव आणि मनोभाव या दोंहोनाही जैविक दृष्ट्या महत्व असते, मनोभावोत्पत्तीच्या वेळी सुखद वा असुखद भाव जाणवत असतो. तथापि, अभिसरण–अपसरणतेस निश्चित व वैशिष्ट्यपूर्ण रूप प्राप्त होते (उदा. आलिंगन, पलायन, आक्रमक वर्तन) ते मनोभावांमुळे होत असते. केवळ सुख-असुख भावांमुळे नव्हे.

Table 3: Irrelevant Result dataset with sample data

Sr. No.	Name of Dataset	Sample Data
1	Que_Ref_WSD_txt	पर्यावरण (Environment) आणि शेती यांच्यातील संबंधांवरील नवीन अभ्यासात, संशोधकांना वायू प्रदूषण आणि कृषी उत्पादकता (Crop Yield) यांच्यातील खोल संबंध आढळला आहे. ते उपग्रह डेटाचे विश्लेषण करण्यात सक्षम झाले आहेत आणि त्यांना आढळले आहे की जर नायट्रोजन ऑक्साइडचे (Nitrogen Oxides) प्रमाण पुरेसे कमी केले तर जगातील अनेक देशांमध्ये पीक उत्पादनात लक्षणीय वाढ होईल. जगाच्या लोकसंख्येसोबत अन्न संकट (Food Crisis) अधिकाधिक आव्हानात्मक होत आहे. हवामान बदल, जमिनीचे प्रदूषण, मातीची धूप इत्यादी समस्या शेती आणि शेतांच्या उत्पादकतेशी संबंधित आहेत. याचा परिणाम सर्व प्रकारच्या कृषी आधारित उपक्रमांवर होत आहे. धान्य पेरणे किंवा त्याचे क्षेत्र वाढवणे हाही उपाय नाही. पण वायू प्रदूषणात (Air Pollution) घट कृषी उत्पादनात फायदेशीर ठरू शकते का? नवीन अभ्यासात वायू प्रदूषण कमी केल्याने शेती उत्पादनाला दीर्घकाळ कसा फायदा होऊ शकतो, तसेच जमीन आणि पैशांची बचत

		कशी होते याचे वर्णन केले आहे. वायू प्रदूषणामुळे (Air Pollution) होणाऱ्या उत्सर्जनातील केवळ एकाच प्रकारच्या प्रदूषणाचे प्रमाण जर जगाने निम्मे केले तर चीनमध्ये (China) हिवाळी पिकांचे उत्पादन 28 टक्क्यांनी वाढेल आणि जगातील दुसऱ्या क्रमांकाचे भाग 10 टक्क्यांनी वाढतील असा अंदाज आहे. या प्रकरणात नायट्रोजन ऑक्साईड (Nitrogen oxides) महत्वाचे आहेत, जे जीवाश्म इंधनाच्या जाळण्याने वाहने आणि उद्योगांमधून सोडले जाणारे विषारी आणि अदृश्य वायू आहेत. वायू प्रदूषणाच्या (Air Pollution) बाबतीत, नायट्रोजन ऑक्साईड (Nitrogen oxides) उत्सर्जन हे जगातील सर्वाधिक वितरित वायु प्रदूषकांपैकी एक आहे. त्यांच्या प्रदूषणामुळे झाडांच्या पानांचे खूप नुकसान होऊन त्यांची वाढ खुंटते, असे सांगितले जाते. नायट्रोजनचे ऑक्साइड ट्रॉपोस्फियरमध्ये ओझोन आणि एरोसोल देखील तयार करतात. त्यामुळे पिकांना योग्य प्रमाणात सूर्यप्रकाश मिळत नाही. याच संशोधनातील अनेक संशोधकांना गेल्या वर्षीच्या अभ्यासात असे आढळून आले आहे की 1999 ते 2019 दरम्यान, ओझोन, कणिक पदार्थ, नायट्रोजन ऑक्साईड (Nitrogen oxides) आणि सल्फर डायऑक्साइडचे प्रमाण घटल्याने अमेरिकेत कॉर्न पीक (Corn Yield) आणि सोयाबीनच्या लागवडीत 20 टक्क्यांनी वाढ झाली होती. त्यामुळे दरवर्षी 5 अब्ज डॉलर्सची पीक बचत झाली. नायट्रोजन डायऑक्साइड स्थानिक पातळीवर देखील सहज मोजता येतो. आणि त्याची थेट पिकांच्या वाढीशी तुलना केली जाऊ शकते. जेव्हा ते वातावरणात उत्सर्जित होते, तेव्हा ते अल्ट्राव्हायोलेट किरणांशी प्रतिक्रिया करते, जे उपग्रह सहजपणे कॅप्चर करू शकतात. संशोधकांचे म्हणणे आहे की, आपण अंतराळातूनही (Space) कृषी उत्पादन (Crop production) मोजू शकतो, त्यामुळे हे वायू शेतीच्या विविध क्षेत्रांवर कसा परिणाम करतात हे जाणून घेण्याची संधी मिळते. नायट्रोजन डाय ऑक्साईडची (Nitrogen oxides) तुलना जगाच्या विविध प्रदेशांतील शेतजमिनीच्या हिरवळीशी करताना संघाला सातत्याने नकारात्मक परिणाम आढळून आले. सर्वाधिक नुकसान चीनमध्ये झाले आहे, त्यासोबतच गव्हासारख्या हिवाळी पिकांचेही नुकसान झाले आहे.

B. Preprocessing

The main aim of information retrieval over the ambiguous word from the relevant and irrelevant data. In this step various action such tokenized, pos tagged, and lemmatized information as can be taken to preprocess the Marathi data.

Tokenization: Tokenization involves breaking down the Marathi text into individual units, usually words and sub-word units. The tokenization helps create a more granular representation of the text, enabling the system to better understand the structure of sentences and phrases.

Part-of-Speech (POS) Tagging: POS tagging assigns grammatical categories such as nouns, verbs, adjectives to each token in the Marathi data. The POS tagging is important for disambiguating words that may have multiple meanings based on their grammatical context. This information aids in understanding the syntactic structure of sentences.

Lemmatization: Lemmatization involves reducing words to their root form. In Marathi data, where words can have various inflections and conjugations, lemmatization helps in standardizing the representation of sentences. This is particularly important for grouping different forms of the same word together, thereby reducing ambiguity and improving the accuracy of query refinement.

Stopword Removal: Removing stopwords can help reduce noise in the Marathi data and improve the efficiency of query analysis. This is particularly important in information retrieval, where irrelevant words may hinder the identification of relevant information.

Normalization: Normalization involves standardizing the text by converting it to lowercase or addressing issues like diacritics and accent marks. This step ensures consistency in the representation of words, reducing the chances of missing relevant information due to case differences.

C. Vector Space Model

The vector space model utilized in this research for Marathi language query refinement encapsulates the essence of document representation and query interpretation. Through the application of TF-IDF vectorization, each query and document are transformed into high-dimensional vectors, where the importance of terms is captured through their frequency and inverse document frequency. This numeric representation allows for the comparison and analysis of queries against a corpus of Marathi documents, enabling the identification of relevant information. By using the TF-IDF scores, ambiguous words in queries are disambiguated, and relevant terms are emphasized, facilitating more accurate information retrieval in the Marathi language.

Query Analysis: Query analysis using TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is a technique in information retrieval to represent and analyze ambiguous words in queries and retrieve relevant information. This approach is used for query refinement over the Marathi language.

TF-IDF Vectorization: TF-IDF vectorization combines TF and IDF to create a numerical representation of a query. Each term in the query is assigned a TF-IDF weight, which reflects its significance in the context of the specific query and the broader Marathi document corpus. This vectorization process transforms the query into a high-dimensional vector, where each dimension corresponds to a unique term, and the values represent the TF-IDF scores. The TF-IDF scores for all terms in a document or query form a vector. Each term corresponds to a dimension in this vector, and the TF-IDF score is the value in that dimension. The resulting vectors are high-dimensional vector, as most terms have low TF-IDF scores in given query.

$$TF\ (t) = \frac{Total\ number\ of\ occurrences\ of\ relevant\ term\ t\ in\ the\ document}{Total\ number\ of\ relevant\ query\ in\ the\ document} \quad (1)$$

The IDF is calculated of a relevant term using equation 2

$$IDF(t) = \log\left(\frac{N}{n_t+1}\right) \quad (2)$$

The TF-IDF score is evaluated using equation 3

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (3)$$

The resulting vectors are high-dimensional and sparse, as most terms have low TF-IDF scores in any given query.

$$TF - IDF\ Vector = [TF - IDF(t1), TF - IDF(t2), \ldots\ldots TF - IDF(tn)] \quad (4)$$

Where, $t1, t2, \ldots, tn$ are the relevant term in the query. This numeric representation captures the importance of each relevant term within the query and is used for information retrieval.

Query Refinement: After analyzing the TF-IDF scores, the proposed query refinement model can identify key terms that is used to retrieve relevant data over the Marathi. This information can be used to refine the query by focusing on to improve retrieval accuracy in Marathi language documents. The Rocchio algorithm is used retrieve the relevance query results into the vector space model. The vector space model is based on three meta heuristics strategies such as term weight, length of document normalization and similarity index.

The term weight shows that term t represents the index for each Marathi document d if it is frequently occurrences in it. Learning happens when create a vector $\bar{v_j}$ for each class V_j by combining Marathi document vectors. First, add up the normalized Marathi document vectors of the relevant result and Irrelevant result for a class. After that, the vector is calculated as a weighted difference between these two sums.

$$\bar{v_j} = (d^{(1)}, \ldots . d^{(F)}) \quad (5)$$

The Marathi data with similar contents have same vectors. Each data shows the different words wi. The $d(i)$ for a data d is measured as an integration of statistical measures $TF(wi,d)$ and $DF(wi)$. The $TF(wi,d)$ represents the frequency of word wi presents in the Marathi data d and the $DF(wi)$ represents the amount of data in each word wi presents at least ones. The $IDF(wi)$ is measured based on the Marathi document frequency.

$$IDF(wi) = \log\left(\frac{|D|}{DF(wi)}\right) \quad (6)$$

Where |D| represents the amount of Marathi data. if a Marathi word appears in many data, its inverse data frequency is low, but if it appears in only one data, its inverse data frequency is high. The so that weight d(i) of Marathi word wi in Marathi dataset d then

$$d(i) = TF(wi,d) . IDF(wi) \quad (7)$$

The Marathi word weight shows that a word wi of query is crucial for index term for marathi dataset d, Whenever the data is occurring many times in the dataset, the TF is high.

To combine the Marathi dataset vector $\bar{v_j}$ into the sample vector for each class V_j. Initially both the normalized Marathi dataset vector of the relevant of class and Irrelevant are combined. The Sample vector is then measured the weight difference of both the class.

$$\bar{v_j} = \alpha \frac{1}{|V_j|} \sum_{\bar{d} \in V_j} \frac{\vec{d}}{||\vec{d}||} - \beta \frac{1}{|D - V_j|} \sum_{\bar{d} \in D - V_j} \frac{\vec{d}}{||\vec{d}||} \quad (8)$$

α and β is the variables the adjust the impact of both the relevant and Irrelevant words. The Rocchio needs the Irrelevant of the vector $\bar{v_j}$ are initially set to 0. Based on the cosine similarity measures and $\alpha = \beta = 1$. Rocchio represents the sample vector that maximizes the mean of similarity of relevant training words using the sample vector V_j subtract the mean of similarity of the Irrelevant training words using the sample vector V_j.

$$\frac{1}{|V_j|} \sum_{\bar{d} \in V_j} \cos(\bar{v_j}, \vec{d}) - \frac{1}{|D - V_j|} \sum_{\bar{d} \in D - V_j} \cos(\bar{v_j}, \vec{d}) \quad (9)$$

The final result set of sample vector, each vector for both classes that shows to trained model. This model can be used to retrieve the relevant Marathi data d". This new data is defined as vector d'. To retrieve the d' the cosine similarity of sample vector $\overline{v_j}$ and $\overrightarrow{d"}$ are measured. The Marathi relevant words d′ is assigned to the class whose sample vector has the highest cosine similarity with it.

$$H_{TFIDF}(d') = \underset{V_j \in v}{\operatorname{argmax}} \cos\left(\overline{v_j}, \vec{d}\right) \quad (10)$$

The equation 10 shows that the relevant class assigned to document d' is determined by finding the class whose average TF-IDF vector is most similar (has the maximum cosine distance) to the TF-IDF vector of document d'. This process helps classify d' into the class that is most similar in terms of the words and their importance (TF-IDF weights) to the Marathi document.

$$H_{TFIDF}(d') = \underset{V_j \in v}{\operatorname{argmax}} \frac{\vec{v}_j}{||\vec{v}_j||} \cdot \frac{\overrightarrow{d'}}{||\overrightarrow{d'}||} \quad (11)$$

$$= \underset{V_j \in v}{\operatorname{argmax}} \frac{\sum_{i=1}^{|F|} v_j^{(i)} . d'^{(i)}}{\sqrt{\sum_{i=1}^{|F|} (v_j^{(i)})^2}} \quad (12)$$

The equation 12 is saying that we are finding the class V_j that maximizes the cosine similarity between its TF-IDF vector and the TF-IDF vector of the document d'. This helps in assigning the document d' to the class that is most similar to it in terms of the words and their importance (TF-IDF weights) to the Marathi document.

Pseudo Code: Query Refinement

Input: Marathi Dataset D1= Query Dataset; D2 = relevant Dataset vectors; D3 = Irrelevant Dataset vectors and the weights α, β, and γ.

α = 1.0 # Weight for the original query
β = 0.75 # Weight for the relevant documents
γ.= 0.25 # Weight for the Irrelevant documents

Output: Classify relevant and Irrelevant Results

Steps

Step-1: Calculate Relevant and Irrelevant Sums

$$\overline{v_j} = \alpha \frac{1}{|V_j|} \sum_{\vec{d} \in V_j} \frac{\vec{d}}{||\vec{d}||} - \beta \frac{1}{|D - V_j|} \sum_{\vec{d} \in D - V_j} \frac{\vec{d}}{||\vec{d}||}$$

Step-2: Calculate Refined Query Vector

$$H_{TFIDF}(d') = \underset{V_j \in v}{\operatorname{argmax}} \frac{\vec{v}_j}{||\vec{v}_j||} \cdot \frac{\overrightarrow{d'}}{||\overrightarrow{d'}||}$$

Step-3: Return Refined Query Vector:

$$Q = \underset{V_j \in v}{\operatorname{argmax}} \frac{\sum_{i=1}^{|F|} v_j^{(i)} . d'^{(i)}}{\sqrt{\sum_{i=1}^{|F|} (v_j^{(i)})^2}}$$

Pseudo Code: *SVMQueryRefinement*

Input:
q: The input Marathi query
$RelDocs$: List of relevant documents
$IrrelDocs$: List of irrelevant documents

Output:
Refined query vector q'

Steps:

1: Convert query and documents to TF-IDF vectors:

$q_v \leftarrow Vectorize(q)$
$RelVectors \leftarrow \{Vectorize(d) \mid d \in RelDocs\}$
$IrrelVectors \leftarrow \{Vectorize(d) \mid d \in IrrelDocs\}$

2: Train an SVM model
$X \leftarrow RelVectors \cup IrrelVectors$
$y \leftarrow \{1 \mid d \in RelDocs\} \cup \{0 \mid d \in IrrelDocs\}$
$model \leftarrow TrainSVM(X, y)$

3: Predict relevance using the SVM model
$relevanceScores \leftarrow model.predict(q_v)$

4: Refine the query based on relevance scores
$DecisionFunctionValues \leftarrow model.decision\_function(RelVectors \cup IrrelVectors)$

5: Calculate weighted sum of relevant and non-relevant documents based on decision function values

$$WeightedRelevantSum \leftarrow \sum_{i=1}^{|RelDocs|} RelVectors[i] \times DecisionFunctionValues[i]$$

$$WeightedIrrelevantSum \leftarrow \sum_{i=1}^{|IrrelDocs|} IrrelVectors[i] \times DecisionFunctionValues[|RelDocs| + i]$$

6: Refine the query vector:

$$q' \leftarrow q_v + (WeightedRelevantSum - WeightedIrrelevantSum)$$

7: Return refined query vector:

$$return\ q'$$

Above pseudocode shows that to Vectorize converts text into TF-IDF vectors, TrainSVM trains an SVM model using relevant and Irrelevant document vectors, and predict calculates relevance scores for the query. The decision_function method is used to obtain the decision function values for relevant and Irrelevant documents. Finally, the query vector is refined based on the weighted sum of relevant and Irrelevant document vectors using the decision function values.

4. RESULT AND DISCUSSION

The experimental model has been performed on Python 3.7 or later, essential libraries like NLTK, SpaCy, Scikit-learn, TensorFlow, PyTorch, and Gensim, along with development environments Google Colab. The hardware requirements include a computer with at least an Intel i7 12th Generation processor, 16 GB RAM, 1TG SSD, and an NVIDIA GEFORCERTX 2070 4 GB GPU, ensuring efficient processing and analysis. The performance of the proposed model has been measured using various parameters such as accuracy, precision, recall and f1-score as follows.

$$Accuracy = \frac{T\_P + T\_N}{T\_P + T\_N + F\_P + F\_N} \quad (39)$$

$$Precision = \frac{T\_P}{T\_P + F\_P} \quad (40)$$

$$Recall = \frac{T\_P}{T\_P + F\_N} \quad (41)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (42)$$

TABLE I. PERFORMANCE ANALYSIS OF SVM CLASSIFIER

Model	Accuracy	Mean Squared Error	Specificity	Sensitivity
SVM	0.79	0.2077	1.0	0.4754

Table 1 presents the performance analysis of an SVM classifier for query refinement in the Marathi language. The classifier achieved an accuracy of 0.79, indicating that it correctly classified documents as relevant and Irrelevant 79% of the time. The mean squared error (MSE) is 0.2077, representing the mean squared variance among the values that were expected and those that were observed; lower values denote superior results. The specificity of 1.0 indicates that all Irrelevant documents were correctly classified as such, while the sensitivity of 0.4754 indicates that 47.54% of relevant documents were correctly classified. The SVM classifier shows strong performance in identifying Irrelevant documents but could be further improved in correctly identifying relevant documents.

TABLE II. CLASSIFICATION REPORT OF SVM CLASSIFIER

	Precision	Recall	F1-Score	Support
Irrelevant	0.68	0.96	0.79	93
Relevant	0.83	0.31	0.45	61
Accuracy			0.70	154
Micro Avg	0.75	0.63	0.62	154
Weighted Avg	0.74	0.70	0.66	154

Table 2 shows the classification report provides a detailed analysis of the SVM classifier's performance for query refinement in the Marathi language. For the "Irrelevant" class, the classifier achieved a precision of 0.68, indicating that 68% of the documents classified as irrelevant were actually irrelevant. The recall for this class is high at 0.96, indicating that 96% of the truly irrelevant documents were correctly classified. The F1-score is 0.79 for the "Irrelevant" class. For the "Relevant" class, the precision is higher at 0.83, indicating that 83% of the documents classified as relevant were actually relevant. However, the recall is lower at 0.31, indicating that only 31% of the truly relevant documents were correctly classified. The F1-score for this class is 0.45. The overall accuracy of the classifier is 0.70, indicating that it correctly classified 70% of the documents. The micro-average F1-score is 0.62, and the weighted average F1-score is 0.66, indicating the overall performance of the classifier across both classes.

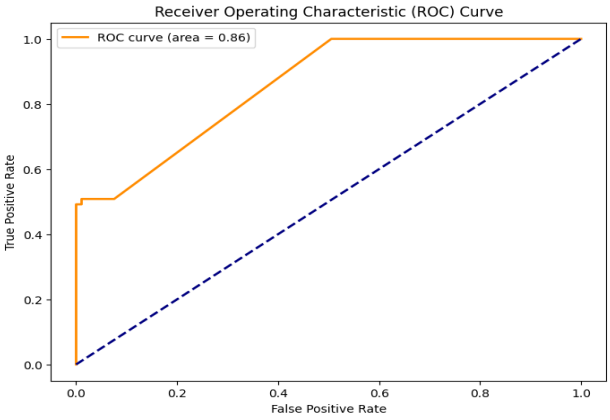

Figure 2: ROC Curve of Proposed Model

An area under the ROC curve (AUC) of 0.86 indicates that the classifier has good performance in distinguishing between relevant and irrelevant documents as shown in figure 2. The AUC value of 0.86 indicates that the classifier is able to correctly rank a randomly chosen relevant document higher than a randomly chosen irrelevant document 86% of the time.

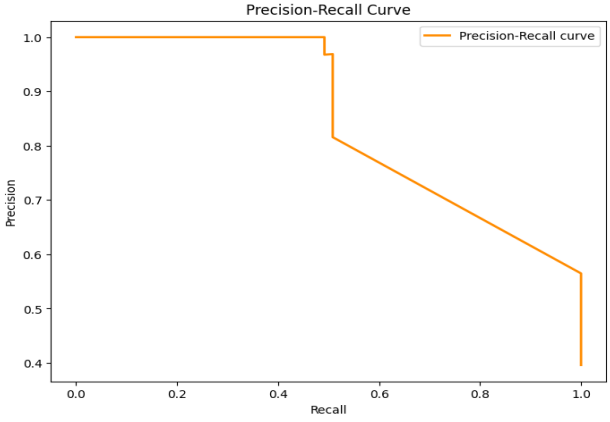

Figure 3: Precision-Recall Curve of Proposed Model

Figure 3 shows the PR curve for query refinement in Marathi language would typically show a curve that starts at the top left corner (high precision) and moves towards the bottom right corner as the threshold is lowered. The area under the PR curve (AUC-PR) is also a measure of the classifier's performance, with higher values indicating better performance. A high AUC-PR indicates that the classifier achieves high precision while maintaining high recall, which is desirable for query refinement tasks where both precision and recall are important.

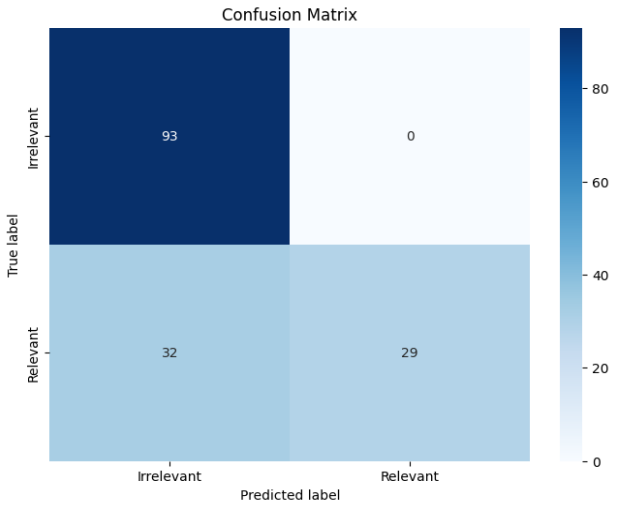

Figure 5: Confusion matrix of SVM Model

Figure 5 shows the confusion matrix to evaluate the performance of a model. TP are the number of documents correctly classified as Irrelevant (93), FP are the number of documents incorrectly classified as Irrelevant (0), FN are the number of documents incorrectly classified as relevant (32), and TN are the number of documents correctly classified as relevant (29).

CONCLUSION AND FUTURE SCOPE

In this paper, proposed model for query refinement in Marathi language information retrieval has shown promising results. The SVM achieved an accuracy score of 0.79, shows to correctly classify documents as relevant or irrelevant. While the classifier demonstrated strong performance in identifying irrelevant documents, there is room for improvement in correctly identifying relevant documents. The classification report further analyzed the classifier's performance, showing high precision and recall for the 'Irrelevant' class, but lower values for the 'Relevant' class. The proposed model shows potential for improving the accuracy score and efficiency of information retrieval in the Marathi language, contributing to the advancement of NLP techniques for less-resourced languages. Further research could focus on refining the proposed framework to handle more complex linguistic nuances and expand its application to diverse domains. Integration with emerging technologies like machine learning and artificial intelligence could enhance the system's performance and adaptability.

REFERENCES

[1] Mosbah, Mawloud. (2023). Query Refinement into Information Retrieval Systems: An Overview. Journal of Information and Organizational Sciences. 47. 133-151. 10.31341/jios.47.1.7.

[2] Liang Yao, Baosong Yang, Haibo Zhang, Weihua Luo, Boxing Chen, (2020) Exploiting Neural Query Translation into Cross

Lingual Information Retrieval, SIGIR eCom'20, arXiv:2010.13659v1 [cs.CL] 26 Oct 2020.

[3] Artem Sokolov, Felix Hieber, and Stefan Riezler. (2014). Learning to translate queries for CLIR. In Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 1179-1182.

[4] Sharma, Sonali, Manoj Diwakar, Prabhishek Singh, Vijendra Singh, Seifedine Kadry, and Jungeun Kim. 2023. "Machine Translation Systems Based on Classical-Statistical-Deep-Learning Approaches" Electronics 12, no. 7: 1716. https://doi.org/10.3390/electronics12071716

[5] Maryamah, Maryamah & Arifin, Agus Zainal & Sarno, Riyanarto. (2019). Query Expansion Based on Wikipedia Word Embedding and BabelNet Method for Searching Arabic Documents. International Journal of Intelligent Engineering and Systems. 12. 202. 10.22266/ijies2019.1031.20.

[6] Litschko, R., Vulić, I., Ponzetto, S.P. et al. On cross-lingual retrieval with multilingual text encoders. Inf Retrieval J 25, 149–183 (2022). https://doi.org/10.1007/s10791-022-09406-x

[7] Sharma, Monika & Morwal, Sudha. (2014). Refinement of search results using cross lingual reference technique. IJARCCE. 8692-8695. 10.17148/IJARCCE.2014.31207.

[8] Stefanescu, Dan & Ion, Radu & Hunsicker, Sabine. (2012). Hybrid Parallel Sentence Mining from Comparable Corpora.

[9] Zeng, Yi & Zhong, Ning & Wang, Yan & Qin, Yulin & Huang, Zhisheng & Zhou, Haiyan & Yao, Yiyu & Harmelen, Frank. (2011). User-centric query refinement and processing using granularity-based strategies. Knowl. Inf. Syst.. 27. 419-450. 10.1007/s10115-010-0298-8.

[10] van Rijsbergen, C.J. (1979) Information retrieval. London, England: Butterworths

[11] Salton, G., and McGill, M.J. (1983) Introduction to modern information retrieval. New York:McGraw-Hill.

[12] Roulland, F. et al. (2007). Query Reformulation and Refinement Using NLP-Based Sentence Clustering. In: Amati, G., Carpineto, C., Romano, G. (eds) Advances in Information Retrieval. ECIR 2007. Lecture Notes in Computer Science, vol 4425. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-71496-5_21

[13] Johannes Leveling and Gareth J.F. Jones, "Sub-Word Indexing and Blind Relevance Feedback for English, Bengali, Hindi, and Marathi IR", in Journal ACM Transactions on Asian Language Information Processing, (TALIP), Vol. 9, No. 3, Article 12, Pub. date: September 2010.

[14] Dong Zhou, Mark Truran, Tim Brailsford, Vincent Wade and Helen Ashman, "Translation Techniques in Cross-Language Information Retrieval", in ACM Computing Surveys, Vol. 45, No. 1, Article 1, Publication date: November 2012.

[15] Saurabh Varshney and Jyoti Bajpai, "Improving Retrieval performance of English-Hindi based Cross-Language Information Retrieval", in Innovation and Technology in Education (MITE), 2013 IEEE International Conference in MOOC, IEEE, pp. 300-305, 20-22 Dec. 2013

[16] Chaware S.M. and Rao S., "Information Retrieval in Multilingual Environment", published in Emerging Trends in Engineering and Technology (ICETET), 2nd International Conference on Emerging Trends in Engineering and Technology, IEEE, pp. 648 – 652, 16-18 Dec. 2009.

[17] Jacques Savoy, Ljiljana Dolamic, and Mitra Akasereh, "Information Retrieval with Hindi, Bengali, and Marathi Languages: Evaluation and Analysis", published in Second International Workshop, FIRE 2010, Gandhinagar, India, pp 334-352 February 19-21, 2010

[18] Jagadeesh Jagarlamudi and A. Kumaran, "Cross-Lingual Information Retrieval System for Indian Languages", published in Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, pp. 80-87, 2008.

[19] Berger, A.; Lafferty, J. Information retrieval as statistical translation. ACM SIGIR Forum. ACM New York, NY, USA, 1999, Vol. 51, pp. 219-226.

[20] Karimzadehgan, M.; Zhai, C. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, 2010, pp. 323-330.

[21] Gao, J.; He, X.; Nie, J.Y. Clickthrough-based translation models for web search: from word models to phrase models. Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 1139-1148.

[22] Karimzadehgan, M.; Zhai, C. Axiomatic analysis of translation language model for information retrieval. European Conference on Information Retrieval. Springer, 2012, pp. 268-280.

[23] Riezler, S.; Liu, Y. Query rewriting using monolingual statistical machine translation. Computational Linguistics 2010, 36, 569-582.

[24] Gao, J.; Nie, J.Y. Towards concept-based translation models using search logs for query expansion. Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 1-10.

[25] Shen, Y.; He, X.; Gao, J.; Deng, L. Latent semantic models with deep neural networks for information retrieval. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval. ACM, 2014, pp. 269-278.

[26] Huang, P.S.; Berard, B.; Gupta, R.; Saini, A.; Saparov, A.; Yin, D. Learning deep structured semantic models for web search using clickthrough data. Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM, 2013, pp. 2333-2338.

[27] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, l.; Polosukhin, I. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, pp. 6000-6010

[28] Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171-4186.

[29] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI 2019.

[30] Zheng, G.; Callan, J. Learning to reweight terms with distributed representations. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp. 575-584.

[31] Zuccon, G.; Koopman, B.; Bruza, P.; Azzopardi, L. Integrating and evaluating neural word embeddings in information retrieval. Proceedings of the 20th Australasian document computing symposium, 2015, pp. 1-8.

[32] Dai, Z.; Callan, J. Context-aware sentence/passage term importance estimation for first stage retrieval. arXiv preprint arXiv:1910.10687 2019.

[33] Frej, J.; Mulhem, P.; Schwab, D.; Chevallet, J.P. Learning term discrimination. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1993–1996.

[34] Clinchant, S.; Perronnin, F. Aggregating continuous word embeddings for information retrieval. Proceedings of the workshop on continuous vector space models and their compositionality, 2013, pp. 100-109.

[35] Vuli´c, I.; Moens, M.F. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp. 363-372.

[36] Kenter, T.; De Rijke, M. Short text similarity with word embeddings. Proceedings of the 24th ACM international on conference on information and knowledge management, 2015, pp. 1411-1420.

[37] Mitra, B.; Nalisnick, E.; Craswell, N.; Caruana, R. A dual embedding space model for document ranking. arXiv preprint arXiv:1602.01137 2016.

[38] Henderson, M.; Al-Rfou, R.; Strope, B.; Sung, Y.H.; Lukács, L.; Guo, R.; Kumar, S.; Miklos, B.; Kurzweil, R. Efficient natural language response suggestion for smart reply. arXiv preprint arXiv:1705.00652 2017.

[39] Gillick, D.; Presta, A.; Tomar, G.S. End-to-end retrieval in continuous space. arXiv preprint arXiv:1811.08008 2018.

[40] [40] Karpukhin, V.; O ˘guz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.t. Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 2020.

[41] Seo, M.; Kwiatkowski, T.; Parikh, A.P.; Farhadi, A.; Hajishirzi, H. Phrase-indexed question answering: A new challenge for scalable document comprehension. arXiv preprint arXiv:1804.07726 2018.

[42] Nie, P.; Zhang, Y.; Geng, X.; Ramamurthy, A.; Song, L.; Jiang, D. Dc-bert: Decoupling question and document for efficient contextual encoding. Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 1829-1832

[43] Ganguly, D.; Roy, D.; Mitra, M.; Jones, G.J. Word embedding based generalized language model for information retrieval. Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp. 795-798.

[44] Roy, D.; Ganguly, D.; Mitra, M.; Jones, G.J. Representing documents and queries as sets of word embedded vectors for information retrieval. arXiv preprint arXiv:1606.07869 2016.

[45] Boytsov, L.; Novak, D.; Malkov, Y.; Nyberg, E. Off the beaten path: Let's replace term-based retrieval with k-nn search. Proceedings of the 25th ACM international on conference on information and knowledge management, 2016, pp. 1099-1108.

[46] Dos Santos, C.; Barbosa, L.; Bogdanova, D.; Zadrozny, B. Learning hybrid representations to retrieve semantically equivalent questions. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Volume 2: Short Papers, 2015, pp. 694-699.

[47] Seo, M.; Lee, J.; Kwiatkowski, T.; Parikh, A.P.; Farhadi, A.; Hajishirzi, H. Real-time open-domain question answering with dense-sparse phrase index. arXiv preprint arXiv:1906.05807 2019.

[48] Lee, J.; Seo, M.; Hajishirzi, H.; Kang, J. Contextualized sparse representations for real-time open-domain question answering. arXiv preprint arXiv:1911.02896 2019.

[49] Luan, Y.; Eisenstein, J.; Toutanova, K.; Collins, M. Sparse, dense, and attentional representations for text retrieval. Transactions of the Association for Computational Linguistics 2021, 9, 329-345.

[50] Gao, L.; Dai, Z.; Chen, T.; Fan, Z.; Van Durme, B.; Callan, J. Complement lexical retrieval model with semantic residual embeddings. European Conference on Information Retrieval. Springer, 2021, pp. 146-160.

[51] Kuzi, S.; Zhang, M.; Li, C.; Bendersky, M.; Najork, M. Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach. arXiv preprint arXiv:2010.01195 2020.