Journal of Information Systems Engineering and Management
2025, 10(13s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Approach to Handle Compound Out of Vocabulary Words in Hindi Web Queries

Amit Asthana[1], Ganesh Chandra[2], Sanjay K. Dwivedi[3]
[1]Department of Computer Science, Babasaheb Bhimrao Ambedkar University, India
[2]Department of Computer Science & Engineering, Madhav Institute of Technology & Science (Deemed University), Gwalior, India
[3]Department of Computer Science, Babasaheb Bhimrao Ambedkar University, India

ARTICLE INFO

Received: 30 Nov 2024

Revised: 16 Jan 2025

Accepted: 30 Jan 2025

ABSTRACT

Introduction: Detection and handling Out of Vocabulary (OOV) words in information retrieval is a challenging task. This problem may become more challenging in case of Cross-Lingual Information Retrieval (CLIR) due to the complications with query translation. Compound Hindi OOV word problem has been less discussed in literature and no appropriate solution has been provided to overcome the issue in CLIR with web queries having such words. These words if not identified, may restrict to understand the proper meaning.

Objectives: The objective of this paper is to understand the impact and to handle web queries involving compound Hindi out of vocabulary words on the retrieval effectiveness.

Methods: This paper proposes an algorithm to detect and handle the compound Hindi OOV words in web queries. The algorithm has been applied on two categories of web query (i.e. having only compound Hindi OOV word and having ateast one such word in it) and the retrieval efficiency is calculated in terms of precision and average precision.

Results: Result highlight an improvement of 8.53% for one-word web queries having only OOV word and 15.68% with queries having more than one word involving at least one OOV word respectively.

Conclusions: The results of the work indicate that the proposed approach detects and handles specific type of OOV words present in Hindi web queries. Improvement in retrieval effectiveness for both types of queries has been observed. Queries having more than one word show better improvements than the queries with single OOV word. This could be attributed to the contextual information offered by the neighboring words.

Keywords: CLIR, Hindi web query, OOV, WX-Notation, compound Hindi OOV

INTRODUCTION

The quality of the documents retrieved through web IR system depends on the effectiveness of the queries. Web queries may sometimes include words and phrases that are not recognized by the search engine. This can lead to poor search results for users.

One of the common issues in web search queries is the presence of OOV words. Words that are not present in the training corpus used for meaning extraction are termed as OOVs, which can be due to misspellings, slang, abbreviations, or other variations of language. These OOV terms may be classified mainly into the categories of named entity (NE), slang, spelling error and foreign term [1]. The presence of these terms may prevent the morphological analyzer from accurately processing the lexical data, affecting meaning extraction for further translation phases. It is therefore essential to detect and handle the OOV words as early as in the morphological analysis phase. The present work uses the queries in Hindi language. It is relatively complex and rich language wherein one or more words can be combined to form a new word using hidden connectors like conjunction and preposition in Hindi, known as समास {samas} or संधि {sandhi} [2].

The two main divisions of natural language are segmented and non-segmented languages. Languages like Chinese, Japanese, etc. where words are not divided by blank spaces are known as non-segmented languages whereas languages like English, Hindi, where words in a sentence are separated by blank spaces, are known as segmented languages. The segmented languages are easier to tokenize in comparison to non-segmented languages because of the available marker (i.e. the blank space). Hindi is a segmented language but various compound-words can be formed by the combination संधि {Sandhi} of two or more words together and these words cannot be separated by blank spaces. A compound word, i.e. a combination of two or more words, not present in a given dictionary or corpus irrespective of the presence of its constituent words is also an OOV word. For example, in English language "dog" and "house" both words are present in a corpus but their combination i.e. "doghouse" is not there then this term is compound OOV. Detection and handling of such compound out-of-vocabulary words in web queries is a more complex task than handling simple OOV words.

Proper feature extraction is very difficult or almost impossible if OOV word(s) present in a sentence. These words are important to handle as early as in the morphological analysis phase. Wrong interpretation for the input may occur if these words are not handled properly.

It is often common for surrounding words to be misidentified if these words are present in a speech recognition system [4]. These errors may spread to later stages of processing, such as translation, comprehension, document retrieval, and term detection, so it will be difficult to obtain the desirable result.

OOV query term may adversely affect the retrieved result in an information retrieval (IR) system. These terms might have been in the input query but gone overlooked, leading to a search term miss. The retrieval damage caused by a miss is determined by characteristics like as query length, query frequency and word frequencies across or within document. Hence, OOV misses will reduce the amount of matching words, affecting retrieval effectiveness.

Table 1. Queries involving compound OOV words and retrieval effectiveness in terms of Precision (P@10) and Average Precision (AP)

Original Query			**Modified Query (Manual)**		
Initial Query	**P @10**	**AP @10**	**Modified Query**	**P @10**	**AP @10**
स्वजनसमुदाय का अर्थ	0.2	0.3	स्वजन समुदाय का अर्थ	0.59	0.62
देशहितकारी योजनायें	0.2	0.49	देश हितकारी योजनायें	0.6	0.6

The potential effects of these terms on IR including web query with OOV words have a direct impact on matches, either by lower scores for matching documents or by not matching at all. Table 1 shows the retrieval effectiveness in terms of precision and average precision for queries involving compound OOV words. It has also been found that sometimes the recognizer generates false in-vocabulary terms by wrongly substituting OOV words, leading to wrong matching with in-vocabulary query terms [14]. This work mainly focuses on detection and handling of compound Hindi OOV words present in web queries.

LITERATURE REVIEW

We could find only a few attempts that have been made to address the issue related to OOV words in Indian languages. Some significant contributions in the related area have been mentioned in this section.

GirishkumarPonkiyaet al. (2018) [3] proposed to paraphrase noun compounds using prepositions by treating noun compounds and their prepositional paraphrases as parallelly aligned word sequences. They used two separate LSTM (Long Short-Term Memory) to encode a noun compound and its prepositional paraphrase, then trained a network to match the encodings of the noun compound and its associated prepositional paraphrase.

Raistrow et al. (2009) [4] divided methods for handling OOVs mainly into two categories: Filler models approach and confidence score approach. The first approach focuses on explicitly modelling OOV words with filler or generic word models. The most frequent method is to utilise a filler or garbage model to absorb OOV words and non- speech objects. This strategy has been successfully employed in key-word spotting, for example, when the recognizer vocabulary predominantly consists of key-words, requiring substantial use of the filler models [5]. The second approach i.e. confidence score method, is a modern technique, focuses on finding OOV words using acoustic scores, statistically generated from the language model and derived from N-best lists as confidence measures.

Girishkumar Ponkiya et al. (2018) [6] presented a dataset for interpreting noun compounds and used Levi's theory [7] that the noun compounds are formed by either predicate deletion or predicate nominalization, to linguistically ground the dataset. They used semantic relation inventory which was based on Frame Net frames and frame elements and conducted these procedures in order to create a standardised dataset with 2,600 samples in the collection. Each noun compound word is annotated with the form of noun compound (predicate deletion vs. predicate nominalization), the frame and frame element through which the noun compound was created in the first place, and its label in three separate datasets.

Patel et al. (2019) [8] used an unsupervised model based on expectation-maximization to induce a transliteration corpus with parallel data, which was then used to train a transliteration model, to show the use of preordering and suffix separation helps in improving the quality of English to Indian language machine translation. Top 100-best transliteration output used for OOV words and were plugged in the translation replacing OOV words and rescored with the language model to get the best translation for source sentence.

Himanshu Choudhary et al. (2020) [9] developed a Neural Machine Translation (NMT) based system for English-Malayalam and English-Tamil language pairs. The OOV issue was addressed using the pre-trained fast text Byte-Pair-Encoded (BPE) and MultiBPE encodings in combination with multi-head self-attention and the BLEU score achieved for the English to Tamil MT system was 24.34, whereas the English to Malayalam MT system of 9.78 respectively.

GirishkumarPonkiya et al. (2020) [10] showed that contextualised language models can be used for unsupervised noun compound paraphrasing and presented two unsupervised techniques for paraphrasing noun compounds: prepositional paraphrasing and free paraphrasing. To produce plausible paraphrases, they used contextualized language models and a feed template and their findings reveal that the unsupervised strategy surpasses supervised approaches for free paraphrasing and are comparable to supervised systems for prepositional paraphrasing.

Jeongin Kim et al. (2021) [11] introduced a Word2VnCR algorithm to replace an OOV word with a semantically related term when an error occurs in morpheme analysis. With the help of this approach, candidate words to be exchanged with the OOV word having the same meaning as OOV are extracted and their semantic similarity to the OOV word's nearby terms can be determined. They also conducted a comparative experiment of Word2VnCR and Word2Vec algorithms and found Word2VnCR to be more efficient in order to replace OOV terms with words that are semantically comparable.

Vijay et al. (2022) [12] proposed a context-based translation algorithm for SMT while CLIR to translate the OOV words using a small bi-lingual parallel corpus and two unlabeled & unrelated raw corpora for both source and target languages, and claimed to reduce the OOV terms up to 0.81% for FIRE 2010 and 1.73% for FIRE 2011.

Baby et al. (2022) [13] suggested a post-processing technique that uses the phone cost function to recover context-based OOV words when the list of context words is known. They also looked at acoustic and phonetic similarity-based cost functions, and their research demonstrates that the suggested approach provides OOV context word recovery of more than 50% across several categories.

Our literature survey reveals that relatively less work has been done in compound OOV words in Indian languages. A study of Hindi compound verbs (i.e. verb + verb) with diagnostic tests for detection as well as automatic axtraction was presented and a morphological analyser using rule-based method for detection and handling of compound words in Marathi has also been developed. However, we could not find such research for compound OOV words in Hindi language.

METHODS

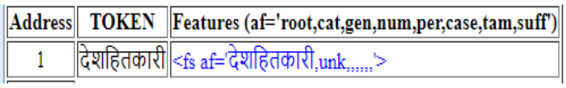
Improving retrieval effectiveness of web queries by handling OOV words is crucial task. One way of identifying these words is to match the tokens with a rich corpus that almost has every possible word which is next to impossible due to continuous expansion in the vocabulary and emergence of the slangs, another way is to handle such terms when they appear [15]. In Hindi language there is a huge possibility for the construction of compound word by the combination of two or more words resulting the presence of OOV. The proposed algorithm focuses on such type of OOV words for handling.

WX-notation [16] has been used in the proposed algorithm for the purpose of transliteration that provides an efficient way to represent terms in understandable and processable format. For example, compound Hindi word देशहितकारी {deshhitkari} {national welfare} and रामभक्त {rambhakt} {devotee of lord Ram} can be represented as 'xeSahiwakArI' and 'rAmaBakwa' respectfully in WX-notation.

We model an algorithm for the purpose of handling OOV compound Hindi words. Morph analyser [17], developed by LTRC (Language Technologies Research Centre) IIIT Hyderabad is used for the purpose of morphological feature (i.e. POS tagging, lexeme, prefix, suffix, etc.) extraction of the words present in the query. The proposed algorithm is given below:

```
Input: I= {user query}
int i=0, j=0, count=0;
String oov, new_word;
while(i<wordCound(I)){
   if (morph analyser identifies I[i])    {
       the word I[i] is not OOV;
   }    else    {
j=0;
       oov=wx_notation_representation_of(I[i]);
       count=length(oov);
       while (j < count)        {
           new_word=new_word+oov[j];
           if (morph analyser identifies new_word)        {
               replace I[i] with new_word;
               new_word=null;
           }
        j=j+1;
 } }
        i=i+1;
 }
```

Let's take an example of user query 'देशहितकारी योजनायें' {deshhitkari yojanayen} as input in the above algorithm. Words of input are analysed one by one with Hindi morph analyser. If the features of the word are extracted, it's not OOV unlike our input where morphological analyser is unable to extract the features of the first word of the query i.e. 'देशहितकारी' {deshhitkari} hence the word is marked as OOV. The output of the analysis for the given word is shown in Fig. 1, where 'unk' represents 'unknown token'.



Address	TOKEN	Features (af='root,cat,gen,num,per,case,tam,suff')
1	देशहितकारी	<fs af='देशहितकारी,unk,,,,,,'>

Fig 1. Outcome of morph analyser for 'देशहितकारी' {deshhitkari} {national welfare}

This OOV word is then converted into WX-notation (i.e. 'देशहितकारी' {deshhitkari} as xeSahiwakArI) and is split letter by letter and checked by the morphological analyser as:

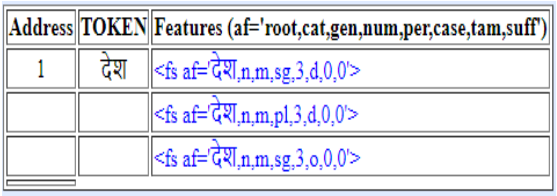
1. new_word = [x] {द्} and oov = [eSahiwakArI] {एशहितकारी}

2. new_word = [xe] {दे} and oov = [SahiwakArI] {शहितकारी}

3. new_word = [xeS] {देश्} and oov = [ahiwakArI] {अहितकारी}

4. new_word = [xeSa] {देश} and oov = [hiwakArI] {हितकारी}

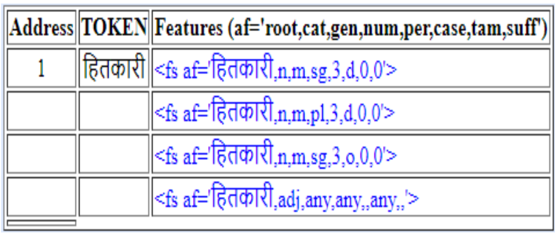
After checking continuously by the morph analyser at each formation of 'new_word', a meaningful word 'xeSa' {देश}{desh}has been detected and its features has been extracted as shown in Fig. 2.



Address	TOKEN	Features (af='root,cat,gen,num,per,case,tam,suff')
1	देश	<fs af='देश,n,m,sg,3,d,0,0'>
		<fs af='देश,n,m,pl,3,d,0,0'>
		<fs af='देश,n,m,sg,3,o,0,0'>

Fig 2. Outcome of morph analyser for 'देश' {desh}

If the new_word is identified by the morph analyser, it is added in place of OOV word in the user query. The new_word is then set to empty and the process continues for the remaining part of the word i.e. 'hiwakArI' {हितकारी}{hitkari}{welfare} and passes to morph analyser as shown in Fig. 3.



Address	TOKEN	Features (af='root,cat,gen,num,per,case,tam,suff')
1	हितकारी	<fs af='हितकारी,n,m,sg,3,d,0,0'>
		<fs af='हितकारी,n,m,pl,3,d,0,0'>
		<fs af='हितकारी,n,m,sg,3,o,0,0'>
		<fs af='हितकारी,adj,any,any,,any,,'>

Fig 3. Outcome of morph analyser for 'हितकारी'{hitkari}{welfare}

As a result of the proposed algorithm the meaningful words 'देश' {desh} {nation} and 'हितकारी' {hitkari}{welfare} are extracted from OOV word 'देशहितकारी' {deshhitkari} and replaced it in the original query to make the query meaningful and system understandable.

RESULTS AND DISCUSSION

Proposed method is motivated by the need to improve retrieval performance by handling the OOV word(s) if present in the query. The experimental analysis has been carried out with queries framed using 50 such words collected from some popular Hindi newspapers and websites. The work mainly focuses on the words that can be split without changing "मात्रा" {maatra}. We took help from the linguists to know the expected result when such queries are subjected to the web search. The precision computation for each query was done through blind evaluation by 10 persons and the average of their findings were taken and analyzed.

Queries are classified into two categories, i.e.

Category-I: One-word query, i.e. the only word in the query is OOV

Category-II: Query having more than one word (with atleast one OOV word).

Retrieved result has been classified as relevant or irrelevant for the purposes of accuracy evaluation based on the values of precision and average precision. It's worth noting that the relevancy of retrieved result is usually a relative measure that varies from one user to the other [18]. The ratio of the number of relevant documents obtained to the total number of documents retrieved is precision.

$$Precision = \frac{Relavant\ Retrieved\ Documents}{Retrieved\ Documents}$$

The Average Precision (AP) is the weighted sum of precisions at each threshold which is a good indicator for comparing how well models organize predictions without taking any specific decision threshold into account.

$$AP = \int P(r)dr$$

Table 2. Results with category-I queries

S. No.	For Original Query			After applying the algorithm		
	Query	Precision (@10)	AP (@10)	Modified Query	Precision (@10)	AP (@10)
1	आसक्तिरहित	0.4	0.58	आसक्ति रहित	0.6	0.81
2	शिवशक्ति	0.5	0.72	शिव शक्ति	0.8	0.89
3	ज्ञानीजन	0.4	0.71	ज्ञानी जन	0.6	0.85
4	शिवधामयात्रा	0.5	0.8	शिव धाम यात्रा	0.7	0.93
5	पंचप्राण	0.7	0.9	पंच प्राण	0.7	0.93
6	कौशल्यानंदन	0.8	0.94	कौशल्या नंदन	0.9	0.97
7	कर्मभूमि	0.4	0.73	कर्म भूमि	0.7	0.92
8	शिवधाम	0.6	0.83	शिव धाम	0.8	0.88
9	सीताराम	0.8	0.94	सीता राम	0.9	0.97
10	आशारहित	0.7	0.93	आशा रहित	0.8	0.96
11	देवात्मा	0.8	0.94	देव आत्मा	0.9	0.97
12	मंदबुद्धि	0.9	0.99	मंद बुद्धि	1	1
13	नित्यवस्तु	0.3	0.42	नित्य वस्तु	0.5	0.55
14	पूर्णविराम	0.7	0.93	पूर्ण विराम	0.8	0.96
15	स्वामीभक्त	0.6	0.88	स्वामी भक्त	0.8	0.96
16	देशहित	0.7	0.93	देश हित	0.9	0.96
17	देशहितकारी	0.7	0.93	देश हितकारी	0.8	0.96
18	कार्यकुशल	0.6	0.86	कार्य कुशल	0.8	0.88
19	दिव्यदृष्टि	0.7	0.93	दिव्य दृष्टि	0.9	0.99
20	सेवाभाव	0.5	0.8	सेवा भाव	0.8	0.96
21	आचारसंहिता	0.6	0.88	आचार संहिता	0.8	0.96
22	लोकसभाध्यक्ष	0.7	0.93	लोकसभा अध्यक्ष	0.8	0.96
23	जलनिगम	0.7	0.93	जल निगम	0.9	0.94
24	दशरथनंदन	0.8	0.92	दशरथ नंदन	0.9	0.95
25	विश्वविख्यात	0.6	0.88	विश्व विख्यात	0.8	0.96
26	विश्वरूप	0.6	0.88	विश्व रूप	0.8	0.96
27	कुंतीपुत्र	0.5	0.82	कुंती पुत्र	0.7	0.93
28	स्वजनसमुदाय	0.2	0.59	स्वजन समुदाय	0.3	0.62
29	शस्त्ररहित	0.6	0.7	शस्त्र रहित	0.8	0.79
30	अविनाशीस्वरूप	0.4	0.73	अविनाशी स्वरूप	0.6	0.83
31	कर्मतत्व	0.7	0.93	कर्म तत्व	0.8	0.96
32	पुण्यात्मा	0.6	0.85	पुण्य आत्मा	0.9	0.97
33	देशाभिमान	0.6	0.87	देश अभिमान	0.8	0.94
34	दर्शनरूप	0.5	0.82	दर्शन रूप	0.8	0.94
35	भोजनगृह	0.7	0.93	भोजन गृह	0.8	0.96
36	जगतुदय	0.2	0.48	जगत उदय	0.3	0.62
37	देवरूप	0.4	0.61	देव रूप	0.5	0.63
38	देशाधिकार	0.3	0.62	देश अधिकार	0.5	0.82
39	पंक्तिबद्ध	0.6	0.88	पंक्ति बद्ध	0.7	0.93

40	अर्धसत्य	0.8	0.96	अर्ध सत्य	0.9	0.99
41	प्रेमास्कत	0.5	0.8	प्रेम आसक्त	0.6	0.85
42	व्योमबाला	0.4	0.73	व्योम बाला	0.7	0.87
43	विद्यादान	0.8	0.96	विद्या दान	0.9	0.99
44	कुलमाता	0.8	0.96	कुल माता	0.9	0.99
45	पंचप्रमाण	0.4	0.73	पंच प्रमाण	0.6	0.83
46	जलसमाधि	0.7	0.93	जल समाधि	0.8	0.96
47	अल्पबुद्धि	0.8	0.95	अल्प बुद्धि	0.9	0.97
48	भवतारण	0.8	0.96	भव तारण	0.9	0.99
49	शिवद्रोह	0.7	0.92	शिव द्रोह	0.8	0.95
50	कुलद्वार	0.1	0.3	कुल द्वार	0.3	0.59
Total		29.4	41.14		37.5	44.95
Mean Average Precision			0.82			0.89

We have found that AP of the web search queries taken with only one OOV word improved from 0.82 to 0.89 after applying the algorithm which shows about 8.53% improvement as shown in table 2.

Table 3. Results with category-II queries

S. No.	Original Query			After applying algorithm		
	Query	**Precision (@10)**	**AP (@10)**	**Modified Query**	**Precision (@10)**	**AP (@10)**
1	आसक्तिरहित जीवन का लक्ष्य	0.6	0.74	आसक्ति रहित जीवन का लक्ष्य	0.8	0.89
2	शिवशक्ति की आराधना	0.7	0.59	शिव शक्ति की आराधना	0.8	0.89
3	ज्ञानीजन की पहचान	0.5	0.69	ज्ञानी जन की पहचान	0.5	0.56
4	शिवधामयात्रा का समय	0.5	0.71	शिव धाम यात्रा का समय	0.5	0.66
5	पंचप्राण का अध्ययन	0.3	0.39	पंच प्राण का अध्ययन	0.3	0.4
6	कौशल्यानंदन का वनवास	0.7	0.92	कौशल्या नंदन का वनवास	0.9	0.99
7	कर्मभूमि का महत्व	0.2	0.25	कर्म भूमि का महत्व	0.3	0.4
8	शिवधाम का महत्व	0.6	0.81	शिव धाम का महत्व	0.7	0.83
9	सीताराम मंदिर	0.9	0.96	सीता राम मंदिर	0.9	0.97
10	आशारहित व्यक्ति के लक्षण	0.2	0.17	आशा रहित व्यक्ति के लक्षण	0.2	0.32
11	देवात्मा का स्वरूप	0.3	0.39	देव आत्मा का स्वरूप	0.4	0.54
12	मंदबुद्धि बालक के लक्षण	0.7	0.85	मंद बुद्धि बालक के लक्षण	0.8	0.89
13	नित्यवस्तु के प्रकार	0.2	0.3	नित्य वस्तु के प्रकार	0.3	0.32
14	पूर्णविराम का प्रयोग	0.6	0.84	पूर्ण विराम का प्रयोग	0.7	0.74
15	स्वामीभक्त के लक्षण	0.2	0.45	स्वामी भक्त के लक्षण	0.5	0.54
16	देशहित के प्रकल्प	0.4	0.68	देश हित के प्रकल्प	0.6	0.63
17	देशहितकारी योजनायें	0.2	0.49	देश हितकारी योजनायें	0.6	0.6
18	कार्यकुशल अधिकारियों की कमी का कारण	0.3	0.39	कार्य कुशल अधिकारियों की कमी का कारण	0.3	0.4
19	कलयुग में दिव्यदृष्टि की प्राप्ति	0.7	0.59	कलयुग में दिव्य दृष्टि की प्राप्ति	0.8	0.89
20	निष्काम सेवाभाव का अर्थ	0.6	0.52	निष्काम सेवा भाव का अर्थ	0.7	0.56
21	आचारसंहिता लागू करने के कारण	0.5	0.71	आचार संहिता लागू करने के कारण	0.6	0.52
22	लोकसभाध्यक्ष के अधिकार	0.4	0.61	लोकसभा अध्यक्ष के अधिकार	0.5	0.63
23	जलनिगम में फर्जी भर्ती की जांच	0.7	0.86	जल निगम में फर्जी भर्ती की जांच	0.8	0.91
24	दशरथनंदन की लीला	0.4	0.64	दशरथनंदन की लीला	0.5	0.52
25	विश्वविख्यात शिवमंदिर	0.7	0.86	विश्वविख्यात शिव मंदिर	0.8	0.91
26	विश्वरूप दर्शन	0.9	0.96	विश्व रूप दर्शन	0.9	0.97

27	कुंतीपुत्र का कौशल	0.3	0.26	कुंती पुत्र का कौशल	0.3	0.52
28	स्वजनसमुदाय	0.2	0.3	स्वजन समुदाय	0.59	0.62
29	शस्त्ररहित सैनिक	0.1	0.29	शस्त्र रहित सैनिक	0.2	0.38
30	अविनाशीस्वरूप तत्व	0.4	0.53	अविनाशी स्वरूप तत्व	0.4	0.62
31	कर्मतत्व का महत्व	0.3	0.5	कर्म तत्व का महत्व	0.4	0.66
32	पुण्यात्मा के दर्शन का फल	0.2	0.3	पुण्य आत्मा के दर्शन का फल	0.3	0.32
33	देशाभिमान की परिकल्पना	0.2	0.21	देश अभिमान की परिकल्पना	0.2	0.3
34	दर्शनरूप का अर्थ	0.2	0.3	दर्शन रूप का अर्थ	0.3	0.32
35	भोजनगृह के नियम	0.2	0.25	भोजन गृह के नियम	0.3	0.39
36	जगतउदय का समय	0.1	0.29	जगत उदय का समय	0.3	0.32
37	देवरूप का दर्शन	0.2	0.25	देव रूप का दर्शन	0.2	0.3
38	देशाधिकार का विषय	0.3	0.26	देश अधिकार का विषय	0.3	0.38
39	पंक्तिबद्ध दीप	0.7	0.86	पंक्ति बद्ध दीप	0.8	0.91
40	अर्धसत्य की भयानकता	0.4	0.28	अर्ध सत्य की भयानकता	0.5	0.48
41	प्रेमासक्त व्यक्ति के लक्षण	0.2	0.44	प्रेम आसक्त व्यक्ति के लक्षण	0.4	0.57
42	व्योमबाला का कार्य	0.1	0.14	व्योम बाला का कार्य	0.1	0.19
43	उत्तमदान विद्यादान	0.8	0.89	उत्तम दान विद्या दान	0.9	0.95
44	कुलमाता को पूजने के पीछे का रहस्य	0.5	0.63	कुल माता को पूजने के पीछे का रहस्य	0.5	0.65
45	पंचप्रमाण का अध्ययन	0.1	0.19	पंच प्रमाण का अध्ययन	0.1	0.29
46	जलसमाधि का अनुभव	0.8	0.89	जल समाधि का अनुभव	0.9	0.94
47	अल्पबुद्धि मनुष्य	0.7	0.79	अल्प बुद्धि मनुष्य	0.7	0.82
48	भवतारण बोध का अर्थ	0.4	0.28	भव तारण बोध का अर्थ	0.5	0.48
49	शिवद्रोह की प्रसांगिकता	0.2	0.17	शिव द्रोह की प्रसांगिकता	0.2	0.4
50	कुलद्वार की सत्यता	0.1	0.14	कुल द्वार की सत्यता	0.1	0.29
Total		20.7	25.81		25.19	29.58
Mean Average Precision			0.51			0.59

Table 3 shows the AP for the original queries with more than one word having at least one OOV word i.e. 0.51 and after applying the algorithm it improves to 0.59, i.e. 15.68% improvement.

We have also found that the improvement in the result for the category-I queries is higher than the category-II query. As per our observations the improvement in the latter type of queries may be attributed by the context provided by the surrounding words.

CONCLUSION

This paper mainly focuses on the compound Hindi OOV words present in the web queries which is still relatively less explored area of IR. Our work demonstrates that the inclusion of OOV terms in web search queries generally results in decreased retrieval efficiency. The approached used in this work is simple and is capable to identify and extract the meaningful terms within the compound Hindi OOV words. Through the proposed method improves Average Precision (AP) by 8.53% for category-I queries and 15.68% for category-II queries, the work can be extended further for the identification of intra-relationship between the terms extracted from such compound words.

REFRENCES

[1] SoHyun Park, Afsaneh Fazly, Annie Lee, Brandon Seibel, Wenjie Zi and Paul Cook. Classifying Out-of-vocabulary Terms in a Domain-Specific Social Media Corpus. https://aclanthology.org/L16-1474.pdf

[2] Utpal Sharma, Jugal K. Kalita, and Rajib K. Das. 2008. Acquisition of Morphology of an Indic Language from Text Corpus. ACM Transactions on Asian Language Information Processing 7, 3, Article 9 (August 2008), 33 pages. DOI: https://doi.org/10.1145/1386869.1386871.

[3] GirishkumarPonkiya, Kevin Patel, Pushpak Bhattacharyya and Girish Palshikar; Treat us like the sequences we are: Prepositional Paraphrasing of Noun Compounds using LSTM, COLING 2018, Santa Fe, New-Mexico, USA, August 20-26, 2018. https://aclanthology.org/C18-1155

[4] A. Rastrow, A. Sethy and B. Ramabhadran, "A new method for OOV detection using hybrid word/fragment system," 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, 2009, pp. 3953-3956, doi: 10.1109/ICASSP.2009.4960493

[5] I. Bazzi, Modeling Out-of-Vocabulary Words for Robust Speech Recognition, Ph.D. thesis, MIT, 2002. http://hdl.handle.net/1721.1/29241

[6] GirishkumarPonkiya, Kevin Patel, Pushpak Bhattacharyya and Girish K. Palshikar, Towards a Standardized Dataset for Noun Compound Interpretation, LREC 2018, Miyazaki, Japan, May 7-12, 2018. https://aclanthology.org/L18-1489

[7] Levi, J. N. (1978). The syntax and semantics of complex nominals. Academic Press New York. https://doi.org/10.2307/412592

[8] Patel, R., Pimpale, P. Sasikumar, M. (2019). Machine Translation in Indian Languages: Challenges and Resolution. Journal of Intelligent Systems. 28(3): 437-445. https://doi.org/10.1515/jisys-2018-0014

[9] Himanshu Choudhary, Shivansh Rao, and Rajesh Rohilla. 2020. Neural Machine Translation for Low-Resourced Indian Languages. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 3610–3615, Marseille, France. European Language Resources Association. https://aclanthology.org/2020.lrec-1.444

[10] GirishkumarPonkiya, Rudra Murthy, Pushpak Bhattacharyya and Girish Palshikar; Looking inside Noun Compounds: Unsupervised Prepositional and Free Paraphrasing, In Findings of Int'l Conf. on Empirical Methods in Natural Language Processing (findings of EMNLP),16-20 November, 2020. https://aclanthology.org/2020.findings-emnlp.386

[11] Kim, Jeongin, Taekeun Hong, and Pankoo Kim. "Replacing out-of-vocabulary words with an appropriate synonym based on Word2VnCR." Mobile Information Systems 2021 (2021). https://doi.org/10.1155/2021/5548426

[12] Vijay Kumar Sharma, Namita Mittal & Ankit Vidyarthi (2022) Context-based Translation for the Out of Vocabulary Words Applied to Hindi-English Cross-Lingual Information Retrieval, IETE Technical Review, 39:2, 276-285, DOI: https://doi.org/10.1080/02564602.2020.1843553

[13] Baby, A., Vinnaitherthan, S., Kerhalkar, A., Jawale, P., Adavanne, S., & Adiga, N. (2022). Context-based out-of-vocabulary word recovery for ASR systems in Indian languages. arXiv preprint arXiv:2206.04305. https://arxiv.org/abs/2206.04305

[14] Vallez, Mari & Pedraza-Jimenez, Rafael. (2007). Natural Language Processing in Textual Information Retrieval and Related Topics [on line]. Hipertext.net. num.5. http://www.hipertext.net.

[15] Bojar, Ondřej & Diatka, Vojtvech & Rychlý, Pavel & Straňák, Pavel & Suchomel, Vít & Tamchyna, Alevs & Zeman, Daniel. (2014). HindEnCorp -- Hindi-English and Hindi-only Corpus for Machine Translation. 3550-3555. https://aclanthology.org/L14-1643/

[16] Akshar Bharati; Vineet Chaitanya; Rajeev Sangal (1996). "Appendix B". Natural Language Processing: A Paninian Perspective (PDF). Prentice-Hall of India. pp. 191–193. ISBN 9788120309210. Retrieved 16 February 2014.

[17] Hindi Morph Analyzer, http://sampark.iiit.ac.in/hindimorph/web/restapi.php/indic/morphclient.

[18] Kumar, Devendra, and Faiyaz Ahamad. "Opinion Extraction using Hybrid Learning Algorithm with Feature Set Optimization Approach." Journal of Electrical Systems 20.3 (2024): 1922-1932.