ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse02942020.pdf
https://doi.org/10.30534/ijatcse/2020/02942020

Impact of Morphology on the performance of Search Engine while retrieving information in Hindi Language

Akanksha Sharma[1], *Vikash Yadav[2]
ABES Engineering College, Ghaziabad, Uttar Pradesh, India
akanksha.gautam.1311@gmail.com[1], vikash.yadav@abes.ac.in[2]

ABSTRACT

The digital world is moving and growing at a tremendous speed in the current time. We are dealing with huge chunk of data that is available on World Wide Web (WWW) while searching for any sort of information. Not only is the data available huge but along with that we see a large proportion of different languages being supported by search engines. In the current time if we take into account Hindi, we have in India around 692 million users speaking this language. This bring us to the concern that we must have more and more of search engines supporting Hindi web searches like we have for other foreign languages so as to meet the needs of natives. But at present the Hindi search of information on web is still a tedious task which needs improvement. The drawbacks these search engines are facing for Hindi information retrieval is due to the complexity we see in the morphology of the language. Not only this we come across word sense disambiguation but polysemy, phonetics, homonymy also comes into picture. The main objective of our paper hence shall be seeing how various search engines perform while searching some predefined queries in Hindi.

Key words: POS Tagging, Natural language Processing, Morphological Ambiguity, Tagging Approaches

1. INTRODUCTION

As Indian users on Internet are growing rapidly, it becomes vital to enhance and upgrade the existing retrieval mode of information so as to make searching a not so tedious task. If we try to understand the criticality of the work, we would find that Hindi is the 3rd most spoken language all across the globe with 544 million speakers that includes 329 million of native speakers, 215 million non-native speakers. Not only is this Hindi also the mother tongue for northern and central states of India. Upgradation of search engines for Hindi information retrieval becomes a quite an important factor as this is the language opted by Indian government for all the official purpose. Hindi now is a primary part of daily newspapers that people read which both are online and offline, a daily dose of entertainment in the form of radio and TV Channels which people enjoy. Seeing all this we find no other reason but to enhance what we have done in the field so far.

Rural areas are seeing a soaring rise of internet users, which accounts for more than 134 million users on internet and with smartphone users browsing internet we have 87 million in count approximately. These users are growing at the rate of 47% per year. This growing crowd of people crave for more and more content on Web for entertainment knowledge skill development and personal engagements. But the demand the users are having for local languages is not meeting up with the sufficient supply. Local search engines came on to the rescue few years back but due to strong market competition from global leaders like Google Yahoo and Bing a few to mention has stopped their growth in this domain. Search engines like Dwaar, Khoj, Guruji and Raftaar showed no success in the market. Guruji was involved in copyrights infringement, Raftaar in spite of several years of operation struggled with market gains. People like Peeyush Bajpai founder of Raftaar back in 2005 realised the need of providing small town and rural areas with the access of local language content through search engines as these were the people who in spite of having access to mobile ,computers net connections were lagging behind. But the downfall yet remained unstoppable for Raftaar. Other than the major factor of global market ruling the WWW, the low penetration of internet in tier II and tier III regions where there was actual need of local language search accounted for its lack of functioning. The other blockers being lack of good Indian content few years back and feasible revenue generation [1].

2. SEARCH ENGINES FOR HINDI INFORMATION RETRIEVAL

2.1 Google

Google is the most popular search engine we know and is used on the World Wide Web having a market share of 92.62%. Results we get on searching is based on prioritizing the content on the basis of Page Rank algorithm. In 2016 Hindi tab was introduced in google for mobile phone and since then, the searching in Hindi Language has raised 10 folds. Not only this giants now have supported seven local languages (Tamil, Telugu, Bengali, Guajarati, Marathi, Kannada, Malayalam) and are about to support three more shortly (Oriya, Urdu, Punjabi) too.

2.2 Yahoo

Yahoo as of date ranks 4th in the world wide search engine list with a market share of 2.32%. In spite of not being a web crawler engine Yahoo was the only first search engine that went popular initially. It indexes web pages that take into account all the formats known (PDF, spreadsheets, word, Excel) which the user is looking for in the result. Having a user of more than *300,000,000* approximately *it* becomes the 3rd most widely used web based search engine in the world.

2.3 Bing

Bing is owned by Microsoft and is a well-known search engine supporting Hindi language now. Getting close to the roots it supports Tamil, Telugu, Bengali, Marathi, Malayalam, Kannada, Gujrati and Punjabi. These languages have been added to text-to-text speech support. As of date Bing stands 3rd as a globally used search engine with a volume of query being 4.58%.

3. FACTORS AFFECT THE INFORMATION RETRIEVAL PERFORMANCE FOR HINDI LANGUAGE ON WEB

Search when it comes to Hindi language on web is associated with a lot of problems. These conditions range from multilingual nature, inflected morphology, spelling with variants, spellings incorrectly spelled and few more to be precise. We shall discuss all of these in details.

3.1 Factors related to morphology

Morphology relates to the linguistics of any language. It is a study of internal structure of the smallest unit of a sentence formation i.e. word. The study and hence this factor becomes quite important because Hindi is rich morphologically like any other Indo-Aryan language. The study helps us to analyse how words are formed from the smallest meaningful units called morphemes or root words. We shall see how the retrieved information varies while we consider the changes in query by first using root word in it and second time avoiding the use of root word. Let us consider few examples of root words along with its variants [2].

- बोल (to speak)-> बोली (speech)

  Here बोल becomes the root word whereas बोली becomes an abstract noun derived from the stem

- विदेश (foreign country) -> विदेशी (foreigner)
- अ + हिंसा (violence) = अहिंसा (non-violence)

Now we shall consider query searching where more clearly we are with words the better result we shall get from the search engine. An observation is performed by taking 20 queries of Hindi language manually created from online Hindi newspapers having 2 subparts each. The first part contains the root word and the second part of the query did not instead it contained the morphological variant of the root word. The performance was measured in terms of its precision. The queries where tested in three different search engines supporting Hindi i.e. Google, Bing and Yahoo. Table 1 illustrates 10 of those queries as an example which we have taken for our analysis to calculate the precision value as what matter the most is quality and not quantity. The precision value was calculated by considering top 10 queries returned as result by the search engine. As shown in Table 1. Seeing Figure 1 the precision value analysed can be concluded by saying that, Google out performs other two search engines in terms of giving exact data as required. When the root word was used in the query the result in terms of number that we got was much greater than in the case of second query where we did not use the root word. The precision at the same time which indicates quality of data was also better in the case of first query i.e. Where we used the root or stem word [3].

Table 1: Precision values by different Search Engine

			Google		Bing		Yahoo	
S. No.	**Query**	**Word Highlighted**	**Doc. Received**	**Precision**	**Doc. Received**	**Precision**	**Doc. Received**	**Precision**
1.1	लोक जीवन की कथा	कथा	28,40,000	0.9	2,02,000	0.4	2,00,000.00	0.5
1.2	लोक जीवन की कथाएं		23,90,000	0.7	95,700.00	0.4	90,200.00	0.6
2.1	भारतियों का धार्मिक झुकाव	धार्मिक	4,95,000	0.3	12,300.00	0.1	12,300.00	0
2.2	भारतियों का धार्मिकता के प्रति झुकाव		1,630.00	0.1	581	0.1	580	0.2
3.1	फूल से बने रंग	फूल	69,70,000	0.8	1,12,000	0.6	11,00,000.00	0.7
3.2	फूलों से बने रंग		20,30,000	0.6	86,100.00	0.6	84,600.00	0.6
4.1	शारीरिक समस्या से राहत	समस्या	6,72,000	0.8	84,900.00	0.9	83,700.00	0.7

4.2	शारीरिक समस्याओं से राहत		5,69,000	0.7	45,000.00	0.5	44,200.00	1
5.1	महिला का आत्मनिर्भर होना क्यों जरूरी है	महिला	60,400.00	0.9	80,300.00	0.5	80,100.00	0.4
5.2	महिलाओं का आत्मनिर्भर होना क्यों जरूरी है		51,900.00	0.9	36,400.00	0.5	36,400.00	0.8
6.1	भारत में नागरिक के मूल अधिकार	नागरिक	6,19,000	1	31	0.9	30	1
6.2	भारत में नागरिकों के मूल अधिकार		5,54,000	1	13	0.9	10	0.9
7.1	भारत में कोरोना का प्रभाव	भारत	67,30,000	0.5	7,93,000	0.8	7,89,000.00	0.7
7.2	भारतीय शहरों में कोरोना का प्रभाव		6,04,000	0.4	96,700.00	0.7	96,500.00	0.7
8.1	आर्थिक परेशानी से छुटकारा	परेशानी	1,76,000	0.9	38,400.00	0.6	39,300.00	0.9
8.2	आर्थिक परेशानियों से छुटकारा		1,75,000	0.9	16,200.00	1	16,200.00	0.9
9.1	सफल छात्र की आदत	छात्र	2,86,000	0.5	13,10,000	0	13,10,000.00	0
9.2	सफल छात्रों की आदतें		1,38,000	0.5	21,100.00	0.2	21,000.00	0.3
10.1	किसान की कहानी	किसान	91,90,000	1	1,33,000	1	1,34,000.00	1
10.2	किसानों की कहानी		85,90,000	1	1,23,000	0.2	1,24,000.00	0.1

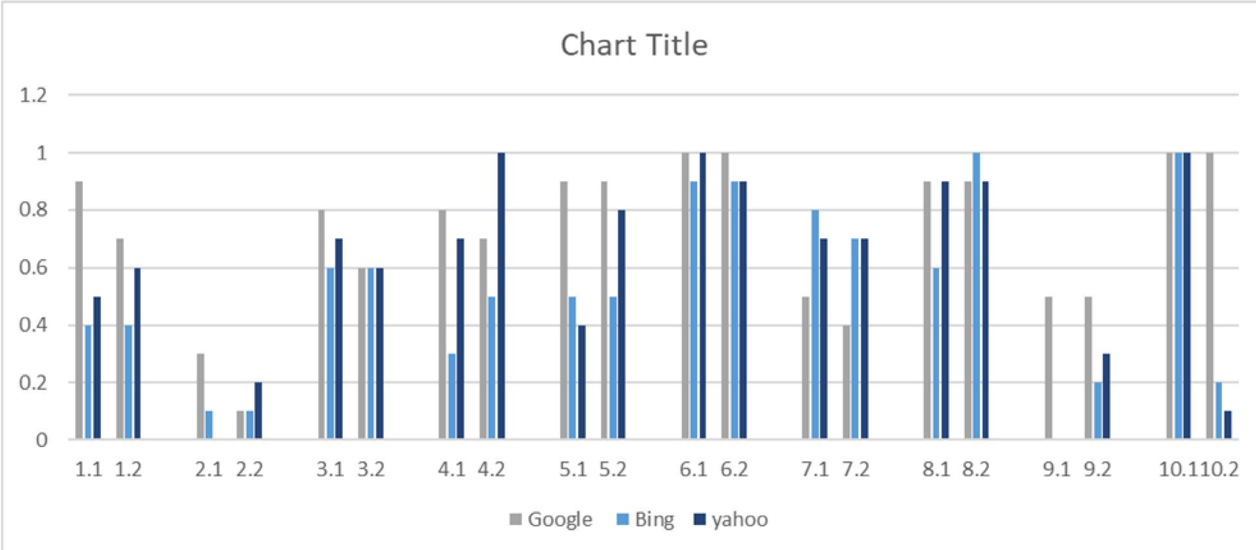

Figure 1: Graphical representation of Precision values by different Search Engine

3.2 Ambiguity in Hindi Word

We shall often come across word in Hindi language having more than one meaning to it. In such a scenario ton find the context of the word in the query becomes a tedious task. These polysemous words most of the time gives us web result that we are not looking for. Let us consider few examples

आम -> फल/ साधारण

Here if we take a note the word it has two meanings at one instance it can mean the fruit mango and in other case normal or common. Again here we shall take 5 random queries to analyse the result in Google, Bing and Yahoo separately to see what result do we get along with its precision value. Table 2 demonstrates the randomly selected queries for observation stating the precision value as obtained from Google similarly Table 2 demonstrates the precision value as obtained from Bing and Table 3 for Yahoo. It was observed that search engines fail to differentiate in many cases as to which context the word is used. We often got mixed results for almost all the queries. For example we take the word "Aam" is some results it meant a commoner whereas for some cases it meant "The AAM Admi Party" and in very few cases "Mango " too.

Table 2: Precision value as obtained from Google Search Engine

Google							
			Result		**Result**		
S.No	**Query**	**Word Highlighted**	**Context 1**	**Total result for context 1**	**Context 2**	**Total Result for context 2**	**Irrelevant Doc.**
1	बाल विकास पर काम	बाल	child	6	hair	1	3
2	कर्म करो फल की इच्छा मत करो	फल	result	9	fruit	0	1
3	सोने की इच्छा	सोने	sleep	2	gold	2	6
4	कुल देवता पूजन	कुल	deity of your house	9	total	0	1
5	आम आदमी का जीवन	आम	commoner	3	mango	0	7

Table 3: Precision value as obtained from Bing Search Engine

Bing							
			Result		**Result**		
S. No	**Query**	**Word Highlighted**	**Context 1**	**Total result for Context 1**	**Context 2**	**Total Result for context 2**	**Irrelevant Doc.**
1	बाल विकास पर काम	बाल	child	4	hair	2	4
2	कर्म करो फल की इच्छा मत करो	फल	result	9	fruit	0	1
3	सोने की इच्छा	सोने	sleep	2	gold	6	2
4	कुल देवता पूजन	कुल	deity of your house	7	total	1	2
5	आम आदमी का जीवन	आम	commoner	6	mango	0	4

Table 4: Precision value as obtained from Yahoo Search Engine

Yahoo							
			Result		**Result**		
S. No	**Query**	**Word Highlighted**	**Context 1**	**Total result for Context 1**	**Context 2**	**Total Result for Context 2**	**Irrelevant Doc.**
1	बाल विकास पर काम	बाल	Child	6	hair	0	3
2	कर्म करो फल की इच्छा मत करो	फल	result	8	fruit	0	2
3	सोने की इच्छा	सोने	Sleep	3	gold	5	2
4	कुल देवता पूजन	कुल	deity of your house	8	total	0	2
5	आम आदमी का जीवन	आम	commoner	6	mango	0	4

3.3 Phonetic Tolerance and variations in spellings for Hindi Language

We all might have encountered in our day to day usage of Hindi Language that Hindi sounds phonetically similar for words written differently with varied range of spellings. These words are used by people interchangeable while searching context on web. These spelling variations are influenced by many factors. To count a few different people in India with different mother tongue, influence of foreign language that we see in Hindi and Phonetic Variation in Hindi and many Indian Languages [4]. In Table 5 are analysis can be concluded stating that no proper rule is laid as to which phonetical variant should be used to gain better result. We tried analysing the variation we get which different variants on Google Bing and Yahoo and in all the three search engines we could find that the result obtained in the form of pages where different in number for every phonetic variant in the entire three browsers [5].

Table 5: Analysis of all three Search Engine

S.No	Query	Phonetically equivalent variants	Number of documents returned			Any Variation Yes/NO
			Google	Bing	Yahoo	
1	शान्त रस का उदाहरण	शान्त रस का उदाहरण	44,400	17,500	17,600	YES
		शांत रस का उदाहरण	1,80,000	26,100	26,000	
2	दाँत खट्टे करना	दाँत खट्टे करना	50,400	16,900	16,900	YES
		दान्त खट्टे करना	1,49,000	14,600	14,600	
3	हिन्दी वर्णमाला का इतिहास	हिन्दी वर्णमाला का इतिहास	2,84,000	21,80,000	22,00,000	YES
		हिंदी वर्णमाला का इतिहास	6,00,000	3,73,000	3,72,000	
4	मुँह फेरना	मुँह फेरना	46,200	5,360	5,350	YES
		मुंह फेरना	76,000	3,120	3,120	
5	आनन्द रामायण	आनन्द रामायण	54,70,000	12,900	13,000	YES
		आनंद रामायण	59,60,000	23,700	23,700	

3.4 Synonymous Words

Words in a sentence as we all know might show multiple attitudes of expressions, having many implications to it and connotations. Choosing the right word synonym for a word while stating a query is essential. This might sound easy is actually a difficult task to handle for both the person querying the data and the data retrieval system. Again an analysis was done taking 5 random queries to see how the information that retrieved was impacted when we used different set synonyms for a word. This analysis was again done of three different search engines i.e. Google, Bing and Yahoo. From the table it is very clear that google is providing more content than the other search engines. It was also observed that every synonymous word returned a different set of data and is being manipulated in a different way. Hence in the case of Hindi it becomes quite essential that we use correct word context while giving query which will obviously come with expertise in the language [6].

Table 6: Superiority of Google over other Search Engine

S.No	Query	Keyword/Synonym	Number of documents returned					
			Google	Precision	Bing	Precision	Yahoo	Precision
1	अतिथि सत्कार	मेहमान	44,700	1.00	3,890	0.40	3,880	0.60
		अभ्यागत	5,580	0.00	597	0.10	597	0.20
		आगन्तुक	21,700	0.10	681	0.00	681	0.00
		पाहूना	463	0.00	20	0.00	20	0.00
		अतिथि	2,53,000	1.00	12,000	0.40	12,000	0.60
2	आँख फड़कना	नैन	3,750	0.00	31	0.00	31	0.00
		आँख	65,400	1.00	7,070	0.80	7,070	0.90
		लोचन	1,140	1.00	17	0.00	17	0.00
3	आकाश में बादल है	आकाश	56,30,000	1.00	1,76,000	1.00	1,76,000	0.90
		अम्बर	15,90,000	0.00	1,25,000	0.90	1,25,000	0.90
4	इच्छा मृत्यु	इच्छा	26,10,000	1.00	26,900	0.80	26,900	0.80
		अभिलाषा	1,39,000	0.30	11,900	0.00	11,900	0.00
5	उम्मीद की किरण	उम्मीद	12,90,000	0.90	2,73,000	0.50	2,73,000	0.50
		आशा	1,02,00,000	0.40	2,73,000	0.40	2,73,000	0.40

4. CONCLUSION

To summarize the whole analysis and conclude we can state that Hindi is a Morphologically Rich Language. The query we give while retrieving information will show varying result if not stated properly. It was observed that When query is supplied with Root word it give better result both in terms of quantity and quality of data. Quality was ensured by calculating the precision. Google out performs the remaining two search engines as it indexes the query. It was further observed that many words that we come across can be ambiguous in nature. This further disturbed the correct retrieval of data. Phonetics plays a major role in Hindi Language. Through our analysis it was clear that none of the search engines were phonetically tolerant. And lastly we saw Synonyms role in query searching. The correct use of word is essential while give any query in any search engine as different synonymous word gave different quantity of content with varying precision.

REFERENCES

[1]https://economictimes.indiatimes.com/tech/internet/strong-competition-from-established-global-players-halts-growth-of-indian-search-engines/articleshow/msid-11634901,curpg-2.cms?from=mdr.

[2] Akanksha Sharma and Vikash Yadav, "Approaches to Part of Speech Tagging in Hindi Language: A Review", International Journal of Advanced Science and Technology (IJAST), ISSN online: 2207-6360 ISSN print: 2005-4238, Vol. 29, No. 5s, pp. 283-291, March 2020.

[3] Kumar Sourabh and Vibhakar Mansotra Performance "Evaluation of Search Engines for Hindi Language Information Retrieval".

[4] GANAPATHIRAJU Madhavi, BALAKRISHNAN Mini, BALAKRISHNAN N. REDDY Raj "Om: One tool for many (Indian) languages" Journal of Zhejiang University SCIENCE ISSN 1009-3095.

[5] S.K. Dwivedi and Parul Rastogi Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow, Uttar Pradesh, India "An Entropy Based Method for Removing Web Query Ambiguity in Hindi Language" Journal of Computer Science 4 (9): 762-767, 2008 ISSN 1549-3636 © 2008 Science Publications. https://doi.org/10.3844/jcssp.2008.762.767.

[6] S.K. Dwivedi and Parul Rastogi Rajesh kr. Gautam "Impact of language morphologies on search engine performance for hindi and English language" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 3, June 2010.