Improving Performance of Search Engines using Morphology

Nargis Parveen, Mohd Athar
Research Scholars
Department of Computer Science, Shri Venkateshwara University
Gajraula, India
nargis.parveen@gmail.com, m.atharjmi@gmail.com

***Abstract*— The morphological, structural and grammar related issues of languages are generally ignored by web searchers during their query formulation and searching. In fact, these factors can be very important in improving the performance of search engines. In this paper, we will show an effort to highlight three factors. Our results show that the performance of the search engines is affected by these factors. The query term ambiguity may sometimes drastically reduce the relevancy of a search engine. Hence it is to be dealt with properly through automated algorithm for disambiguation. Overall, the search engines can be made more users friendly and productive by appropriately handling these issues.**

Index Terms— **Web Ambiguity, Search Engine, Morphology.**

I. Introduction

The term 'morphology' refers to the study of the internal structure of words, and of the systematic form-meaning correspondences between words. Morphology is the study of the structure of words. The structure of words can also be studied to show how the meaning of a given morpheme, or its relation to the rest of the word, varies from one complex word to another. Consider how sun works in the following words: sunbeam, sunburn, sundial, sunflower, sunglasses, sunlight, sunrise, and sun-spot (scientific sense), and sun-spot (tourist sense), and suntan. Inflection does not really yield "new" words, but alters the form of existing ones for specific reasons of grammar. Derivation, on the other hand, does lead to the creation of new words.

Morphology is the field of linguistic which studies word structure and formation [12]. It is composed of inflectional morphology and derivational morphology [13, 14]. Inflection is defined as the use of morphological methods to form inflectional word forms from a lexeme. Inflectional word forms indicate grammatical relations between words. Derivational morphology is concerned with the derivation of new words from other words using derivational affixes. Compounding is another method to form new words. A compound word (or a compound) is defined as a word formed from two or more words written together. The component words are themselves independent words (free morphemes).

A morpheme is a smallest unit of a language which has a meaning. Morphemes are classified into free morphemes and bound morphemes [14, 15]. Free morphemes appear as independent words. e.g. In English, {red}, {house} and {when} are free morphemes. Bound morphemes do not constitute independent words, but are attached to other morphemes or words. Bound morphemes are also called affixes. Morphological structure of English language has a great impact on the performance of the search engines. In this study we have focused on three factors of language morphology that can change or modify a web query i.e. query with root word, query with different synonym and query with various senses.

Methods of Evaluation of Search Engines

There are following methods which are used for the evaluation of search engine:

Precision (P): is the fraction of retrieval documents that are relevant. A high precision means that everything returned was a relevant result, but one might not have found all the relevant items (which would imply low recall). There are variations in the ways of the precision is calculated. TREC almost always uses binary relevance judgments-"either a document is relevant to a query or it is not" [16]. Chu & Rosenthal [17] used a three-level relevance score (relevant, somewhat relevant, and irrelevant) while Gordon and Pathak [18] used a four-level relevance judgment (highly relevant, somewhat relevant, somewhat irrelevant, and highly irrelevant).

RECALL (R): It is the fraction of relevant documents that are retrieved. A high recall means we haven't missed anything but we may have a lot of useless results to sift through (which would imply low precision). But Recall is a difficult measure to calculate because it requires the knowledge of the total number of relevant items in the collection. Chu & Rosenthal's Web search engine study omitted recall as an evaluation measure because they consider it "impossible to assume how many relevant items are there for a particular query in the huge and ever changing Web systems" [17]. Based on the documents retrieved by a search engine (relevant, non relevant), Table 1 below shows the method of computations of precision and recall.

TABLE 1: Precision and Recall Computation Table

	Relevant	**Non-relevant**
Retrieved	True positives (tp) - Correct result	False positives (fp)- Unexpected result
Not retrieved	False negatives (fn) - Missing result	True negatives (tn) - Correct absence of result

The precision and recall can be calculated by the formula shown below: Precision = tp/ (tp+fp)

Recall = tp/ (tp+fn)

Where tp is retrieved relevant result, and fp is retrieved non relevant result, and fn is missing result (i.e. relevant but not retrieved)

Mean Average Precision (Map):

Most standard among the TREC community is Mean Average Precision (MAP), which provides a single-figure measure of quality across recall levels. Among evaluation measures, MAP has been shown to have especially good discrimination and stability. For a single information need, Average Precision is the average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved, and this value is then averaged over information needs.

MAP = Average Precision/ No. of queries

When a relevant document is not retrieved at all, the precision value in the above equation is taken to be 0. Why these methods are used?

These methods are used because the users always want see some documents, and can be assumed to have a certain tolerance for seeing some false positives providing that they get some useful information. The measure of precision and recall concentrate the evaluation on the return of true positive, asking what percentage of the relevant documents have been found and how many false positive have also been returned.

Evaluation Methodology

The U.S. National Institute of Standards and Technology (NIST) have run a large IR test based evaluation series since 1992. Within this framework, there have been many tracks over a range of different test collections, but the best known test collections are the ones used for the TREC Ad Hoc track during the first eight TREC evaluations between 1992 and 1999. TRECs 6 through 8 provide 150 information needs over about 528,000 newswire and Foreign Broadcast Information Service articles. In this work, we have framed the queries based on the TREC pattern and also from the web search engine's log. So our set of test queries used for the evaluation of search engines in this study have a good mix of standard TREC queries and actual user queries from the search engine's log.

Human Relevance Judgments:

It is one of the important issues in performance evaluation of search engines is that whenever human relevance judgment is used, there is a variation in who makes the judgments. TREC leaves relevance judgments to experts or to a panel of experts (Voorchees & Harman, 2001) [16]. However some other researchers (e.g. Chu and Rosenthal, 1996) used human relevance judgment made by researchers themselves. Gordon and Pathak [18] emphasized that relevance judgments can only be made by individual with the original information need. In this study, the human relevance judgments have been done using a mix of the approaches followed by Voorchees et.al (2001) and Chu et.al. (1996).

Precision:

There are variations in the ways how precision is calculated. In this study, the precision is calculated on the binary relevance judgment approach followed by TREC -"either a document is relevant to a query or it is not" [16].

Recall:

Chu & Rosenthal's [17] Web search engine study omitted recall as an evaluation measure because they consider it "impossible to assume how many relevant items there are for a particular query in the huge and ever changing Web systems". In this study too we have omitted the recall as an evaluation measure for the similar reasons.

The computation of precision has been done as follows: Suppose an IR system returns 8 relevant documents and 10 non-relevant documents. There are a total of 20 relevant documents in the collection.

tp (true positive) = 8

fp (false positive) = 10

fn (false negative) = 20-8=12

Precision = tp/ (tp+fp) = 8/ (8+10)

= 8/18 =0.44

Average Precision = sum of all precision/ No. of queries

Mean Average Precision = av. precision/ No. of queries

Factors Affecting Performance of Search Engines

The information retrieval on the web in any language faces numerous challenges. Besides all the technical factors the grammatical and morphological structure of the language is one of the critical factors that can affect the performance of the information retrieval system on the web.

Root Word of the Keywords:

In English prefixes and suffixes (collectively called affixes) are normally used (e.g. s, es, dis, ness, ing etc.) with morpheme (root word) and new words are constructed. These new words are called morphological variants of the stem.

For ex.: increase + ing = increasing, or dis + able = disable. Or happy + ness = happiness.

While searching on the web the query terms given by the users may not be in root form. As there is no restriction/help about how to choose or select the query term, same query may be formed with different morphological variations of its terms. This may lead to variation of results and the relevancy of results by search engines. To analyze this, we took a real time test of Google search engine using a set of 20 web queries (as per the discussion in the previous section). These queries are listed in table 2, and to properly analyze the result each query has been written twice - with root words and without root words.

TABLE 2: Test Query Set For Root Word Analysis

Query with root word	**Query without root word**
Civil Service **exam**	Civil service exam**ination**
Mercury level in **bird**	Mercury levels in bird**s**
water **waste** in India	water wast**age** in India
Fund and grants institution	Fund**ing** and grants institution
beds sharing with **children**	beds sharing with children**'s**

mercury levels is **increase**	mercury levels is increas**ing**
The temperature is **decrease**	The temperature is decreas**ing**
Native **language** of India	Native language**s** of India
merit of democracy	merit**s** of democracy
Use of computer	Use**s** of computer
demerit of democracy	demerit**s** of democracy
advantage of mobile phones	advantage**s** of mobile phones
disadvantage of mobile phones	disadvantage**s** of mobile phones
Imagine power	Imagin**ation** power
power of **battery**	power of batter**ies**
liberty of information act forms	libert**ies** of information act forms
Game is **begin**	Game is begin**ning**
Choose right path	Choos**ing** the right path
Problem is **examine**	Problem is examine**d**
English **query**	English quer**ies**

We then performed Google test for each pair of query set (table 2) and precision values are computed as shown below in the Tables 3 & 4.

TABLE 3. Precision Computation for Queries with Root Words on Google (Using Table 2)

Query	**Doc. Retrieved**	**Precision @10**
1.1	5,350,000	0.55
2.1	35,100,000	0.57
3.1	71,800,000	0.5
4.1	114,000,000	0.66
5.1	25,500,000	0.66
6.1	68,900,000	0.77
7.1	125,000,000	0.77
8.1	5,990,000	0.37
9.1	17,800,000	0.88
10.1	2,900,000,000	0.66
11.1	369,000	0.66
12.1	112,000,000	0.62
13.1	1,270,000	0.88
14.1	126,000,000	0.66
15.1	572,000,000	0.5
16.1	18,700,000	0.6
17.1	17,456,000	0.62
18.1	18,187,000	0.7
19.1	26,432,000	0.57
20.1	9,876,000	0.66
Mean Average Precision = 0.643		

TABLE 4. Precision Computation for Queries without Root Words on Google (USING TABLE 2)

Query	Doc. Retrieved	Precision
1.3	7,920,000	0.55
2.2	26,000,000	0.55
3.2	162,000	0.44
4.3	94,300,000	0.44
5.2	27,200,000	0.57
6.3	68,400,000	0.62
7.3	26,300,000	0.44
8.2	2,780,000	0.77
9.2	7,870,000	0.37
10.2	572,000,000	
11.2	194,000	0.77
12.2	10,200,000	0.44
13.2	1,980,000	0.62
14.2	112,000,000	0.55
15.2	556,000,000	0.55
16.2	15,600,000	0.5
17.2	12,768,000	0.55
18.2	13,145,000	0.6
19.2	23,564,000	0.44
20.2	7,956,000	0.57
Mean Average Precision = 0.5445		

From the Tables 3 & 4, it is clear that when queries are in root form, search engine generally indexes more documents (comparing columns II of tables 3 & 4) i.e. the documents Retrieved are higher. The mean average precision for the root word queries is also higher. It shows that the root word queries are better understood by the Search Engines.

Synonimity:

It is the common characteristics of most of the natural languages. A query term can have a number of representations by its synonym. We observed while working on English language search engines that any word can express a myriad of implications, connotations, and attitudes in addition to its basic ‘dictionary’ meaning. Choosing the right word can be difficult for people. In order to justify this impact of varying synonyms on the web search results, we selected another 20 query set with the help of web query logs. The table 5 below shows the set of queries, where each query been regenerated with a synonyms for one of the terms of query (in bold). The queries of table 5 are examined on the Google search engine and precision is computed for each query it is shown in table 6 & 7.

TABLE5: Test Query set for synonimity word analysis

Original query	**Query with synonyms**
School bus safety	School bus security
Aircraft protection act 2004	Aircraft security act 2004

beds sharing with children	beds sharing with kids
freedoms of information act forms	liberty of information act forms
Aim of project	Objective of project
Top beautiful actress in bollywood	Top gorgeous actress in bollybood
application of internet	uses of internet
Advantage of computer	Merits of Computer
Disadvantage of Computer	Demerits of Computer
Ganga is a large river.	Ganga is a big river.
Atom is made up of tiny particles	Atom is made up of small particles
This is correct answer	This is right answer
Game is start.	Game is begin
This answer is wrong	This answer is false
feel very sleepy	feel very tired
shut the door	close the door
house of rabbit	home of rabbit
difficult problems of algebra	hard problems of algebra
images of caps	images of hats
birthday gift	birthday present

TABLE 6. Precision Computation for Synonimy Using Google (Using Table 5)

Query	Doc. Retrieved	Precision @10
1.1	57,200,000	0.44
2.1	2,090,000	0.77
3.1	4,180,000	0.62
4.1	66,100,000	0.66
5.1	671,000,000	0.66
6.1	5,980,000	0.55
7.1	1,300,000,000	0.5
8.1	11,400,000	0.66
9.1	12,345.000	0.88
10.1	622,800,000	0.77
11.1	564,000,000	0.44
12.1	145,000,000	0.75
13.1	786,000,000	0.66
14.1	111,498,000	0.77
15.1	15,700,000	0.66
16.1	96,000,000	0.44
17.1	14,567,000	0.55
18.1	25,453,000	0.77
19.1	45,600,000	0.77
20.1	15,675,000	0.55
Mean Average Precision = 0.6435		

TABLE 7. Precision Computation for Synonimy Using Google (Using Table 5)

Query	Doc. Retrieved	Precision
1.2	53,800,000	0.66
2.2	40,600,000	0.44
3.2	3,810,000	0.75
4.2	8,040,000	0.44
5.4	662,000,000	0.87
6.2	1,150,000	0.75
7.2	572,000,000	0.55
8.2	98,000,000	0.37
9.2	12,234,000	0.62
10.2	655,700,000	0.55
11.2	675,830,000	0.66
12.2	123,112,000	0.62
13.2	657,000,000	0.62
14.2	104,781,000	0.55
15.2	18,654,000	0.77
16.2	87,678,000	0.33
17.2	12,124,000	0.44
18.2	23,675,000	0.55
19.2	44,134,000	0.66
20.2	16,786000	0.66
Mean Average Precision = 0.593		

The comparative results of the two tables (Table 6 & 7) clearly indicate that search engine (Google) did not properly understand the 'synonym' of a query term. That is why its indexing of documents varies in large number on changing the synonym of a query term. The precision values of the corresponding columns (for one query) of two tables also show variations. This would certainly have an impact on search engine's performance. Our results, however, do not show any trend as to which particular synonym of a query may retrieve more documents and/or higher relevancy.

Sense Ambiguity (Ambiguous Keywords):

Many words are polysemous in nature that is they have multiple possible meaning and senses. Finding the correct sense of the words in the given context is an intricate task. Various researchers (especially Eric Brill [19] and Argaw [20], Navigili and Christopher Stoke [21] and John Tait [22]) have justified the role of Word Sense Disambiguation in the improvement of performance of web searching for English and other languages. Ambiguous keywords deflate the relevancy of the results. We considered 20 queries (based on our discussion in para III) which are normally ambiguous in nature (a query has been considered ambiguous if one of the term of query is ambiguous). Further, in order to analyze the impact of ambiguity over search engine's performance we have tried to manually disambiguate each query with the help of Word Net Database and the search engine in consideration and have shown the effect of ambiguity on the performance of the search engines. This is shown in table 8 where the left side column has query with ambiguity and right side column has manually redesigned query without ambiguity same query

Table: 8: Test Query Set For Ambiguity Analysis

Query with ambiguous word (in bold)	Query with unambiguous words
Wall paint is blue	Wall color is blue
The train is standing on the platform	The train is standing on the railway
There are four seasons in a year	There are four cycle in a year
critical case	critical situation
A bug terminates a program	A error terminates a program
Python are found mostly in rainy	Python snakes are found mostly in
Draw the figure of a flower	Draw the diagram of a flower
Close the door	Shut the door
There should be a break between two	There should be a gap between two
The river is dry	The river is empty
Score of team India in World cup	Run of team India in World cup
balance in my phone	money in my phone
live in present	live in today
aim of a doctor	duty of a doctor
the pitch of sound is high	the level of sound is high
Use of cosine function	Use of cosine expression
The chair of conference	The head of conference
Exercise is necessary to keep our	Physical Exercise is necessary to keep
interest in science	favorite is science
major accident	big accident

The above queries are examined on the Google search engine and the results are shown below in the Tables 9 & 10.

TABLE 9. Precision Computation for Ambiguity Using Google (Using Table 8)

Query	Doc. Retrieved	Precision @10
1.1	140,000,000	0.44
2.1	31,600,000	0.66
3.1	2,860,000	0.37
4.1	175,000,000	0.55
5.1	2,550,000	0.5
6.1	1,020,000,000	0.55
7.1	18,400,000	0.66
8.1	435,000,000	0.33
9.1	2,210,000	0.75
10.1	662,000,000	0.37
11.1	4,420,000	0.22
12.1	325,000	0.44
13.1	12,600,000	0.62
14.1	9,260,000,000	0.44
15.1	16,200,000	0.5
16.1	338,000,000	0.55
17.1	174,000,000	0.66
18.1	335,000,000	0.55
19.1	45,100,000	0.44
20.1	683,000,000	0.75
Mean average precision = 0.5175		

TABLE 10. Precision Computation for Ambiguity Using Google(Using Table 8)

Query	Doc. Retrieved	Precision @10
1.2	374,000,000	0.33
2.2	187,000,000	0.77
3.2	3,150,000	0.44
4.2	374,000,000	0.33
5.2	95,000,000	0.44
6.2	363,000,000	0.55
7.2	66,000,000	0.37
8.2	78,998,000	0.75
9.2	123,000,000	0.44
10.2	112,342,000	0.87
11.2	145,000,000	0.75
12.2	786,000,000	0.66
13.2	111,498,000	0.77
14.2	15,700,000	0.66
15.2	27,200,000	0.57
16.2	68,400,000	0.62
17.2	26,300,000	0.44
18.2	2,780,000	0.77
19.2	572,000,000	0.5
20.2	18,700,000	0.6
Mean average precision = 0.5815		

After examining and comparing the precision values of each queries (Tables 9 & 10), we found that after manual disambiguation of the queries, the precision of 13 out of the 20 queries has improved. The mean average precision has also improved. This shows that the ambiguity in web query can result in poor relevancy of results. Sometimes ambiguity in queries produces adverse results.

II. Conclusion and Limitations

We have evaluated the performance of the English language search engines in the light of their morphological structures and sense ambiguity.

Our results conclude that the performance of the search engines is quite affected by the morphological issues as well as sense ambiguity problems. Ambiguity is the well known problem of the information retrieval setup. Measures are taken to avoid this problem as it affects the relevancy of the results to a great extent.

In the case of web information retrieval the results of queries vary because web is dynamic in nature. Sometimes the ambiguous query may result out in the relevant results and at another time the similar query may result out in the low relevancy results. Therefore the need of ambiguity detection arises, as automatic disambiguation may lead to the wastage of computational power. Hence detection prior to disambiguation is necessary and it is quite evident from the results.

The sense ambiguity problem much affects the search engine performance because the search engines are not capable to cope up this problem. Therefore, to resolve this problem there is a need of Word sense disambiguation (WSD) algorithm. This WSD algorithm is used to disambiguate the sense of the ambiguous words and to improve the search engine performance. But before applying the WSD algorithm the ambiguity detection is necessary. It divides the queries in two parts: ambiguous and unambiguous queries. In our thesis, we have design an algorithm to detect the ambiguity in the query. After this the WSD algorithm is applied only on those queries which are ambiguous. This will increase the performance of search engines. By using the WSD methods we develop an algorithm to resolve the ambiguity from the ambiguous queries.

Acknowledgment

This paper could not be written to its fullest without the guidance of Dr. Avdhesh Gupta, who served as my supervisor, as well as one who challenged and encouraged me throughout my time spent studying under him. He would have never accepted anything less than my best efforts, and for that, I thank him.

References

[1] Sanderson, M., (1994); “Word Sense Disambiguation and Information Retrieval”, Proceedings of SIGIR-94,17th International Conference on Research and Development in Information Retrieval, Dublin, pp. 49-57.

[2] Krovetz, R; Croft, W. B. “Lexical Ambiguity and Information Retrieval” in ACM Transactions on Information Retrieval Systems, Vol. 10(2), Pp 115 –141, 1992

[3] Stokoe, C.M. and Jhon, Tait. (2002); “Automated Word Sense disambiguation for Internet Information Retrieval”. TREC-2002-WEBTRACK

[4] Yarowsky, D. “One Sense Per Collocation” In Proceedings of the ARPA Human Language Technology Workshop, Pp 266 – 271, Princeton, NJ, 1993.

[5] Sanderson, M. “Retrieving with Good Sense” In Information Retrieval, Vol. 2(1), Pp 49 – 69, 2000.

[6] Bybee, J.L. Morphology: a study of the relation between

[7] Dwivedi SK, Rastogi P, Goutam R. Impact of language morphologies on Search Engines Performance for Hindi and English language, IJCSIS, Vol. 8, No.3, 2010.

[8] Voorhees, E.M., & Harman, D. (2001). Overview of TREC 2001. NIST Special Publication 500-250: The 10th text retrieval conference (TREC 2001) (pp. 1-15). Retrieved 17 December 2002 from http://trec.nist.gov/pubs/trec10/papers/overview_10.pdf.

[9] E. Bril and S. Vassilvitskii, “Using WebGraph Distance for Relevance Feedback in Web Search” in Proceedings of SIGIR’06, Seattle, Washington, USA, pp. -147-153, 2006.

[10] A. A. Argaw, “Amharic-English Information Retrieval with Pseudo Relevance Feedback”, in Proceedings of 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, pp. 119-126, 2007

[11] R. Navigili and P. Velardi, “Structural Semantic Interconnection: a knowledge-based approach to Word Sense Disambiguation”, in Journal of Pattern Analysis and Machine Intelligence, Volume 27, Issue 7, pp. 1075 – 1086, July 2005

[12] C. Stokoe and J. Tait, “Towards a Sense Based Document Representation for Internet Information Retrieval”, in Proceedings of SIGIR’03, July 28- August 1, Toronto, Canada, pp. 791-795, 2003

[13] Dwivedi SK, Rastogi P. An Entropy based method for removing web query ambiguity in Hindi language, Journal of Computer Science 4(9): 762-767, 2008 ISSN 1549-3636.

[14] Miguel Angel Rios Gaona, 2009. Web-Base Variant of the Lesk approach to word sense disambiguation. IEEE DOI 10.1109/MICAI.2009.41.

[15] [34]Roberto Navigli and Giuseppe Crisafulli, 2010. Inducing Word Senses to Improve Web Search Result Clustering. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 116-216, MIT, Massachusetts, USA, 9-11 October2010.