F 1 Score Analysis of Search Engines

Sense Retrieval Efficiency of Google, Bing & Yahoo for Single Keyword Queries

Saurabh Ranjan Srivastava[1], Girdhari Singh[2]
Department of Computer Engineering
[1]Swami Keshvanand Institute of Technology Management & Gramothan, Jaipur,
[2]Malviya National Institute of Technology, Jaipur
Email- *[1]saurabh.ranjan.srivastava@gmail.com*
Received 29 August 2016, received in revised form 16 September 2016, accepted 16 September 2016

Abstract: **Word sense disambiguation (WSD), is an open problem of information retrieval domain. WSD leads to retrieval of multiple meanings for a single searched keyword. The efficiency of search results for an automated search machine depends on its efficiency to handle WSD to a large extent. This efficiency depends on the ability to precisely recall the required results against a user query. The weighted harmonic mean of precision and recall, the F-measure, also known as the F1-score is a scale of testing accuracy for an input dataset.**

This paper evaluates the WSD handling capacity of three major search engines, Google, Bing and Yahoo based on their F1 scores. The F1 scores for each search engine are based on their precision and query classification performances [1]. Ten queries are constructed from single keywords that produce different ambiguous meanings for different contexts. With varying index sizes, number of results per page and different search ranking strategies, these search engines responded differently for our set of single keyword queries.

Finally, a comparison of the relevance of search results of Google, Bing and Yahoo for different ambiguous sensed results of single keyword queries is also discussed.

Keywords: **Ranking, relevancy, query, context, ambiguous..**

1. INTRODUCTION

Automated search engines employ search ranking algorithms to return matches of varying quality and precision in their search results [2]. Besides automatically fetched results, paid inclusion of links also increases the complexity of search results. Modern large scale search engines employ hypertext architecture of websites to provide higher quality search results [3]. In this scenario, the major challenge for search engines remains in generating the most relevant search results for a user query by examining the underlying structure of web documents.

Existence of multiple ambiguous meanings for single keyword queries, further elevates this challenge. This paper examines the performance of three search engines in handling the ambiguity of results through word sense disambiguation technique.

2. WORD SENSE DISAMBIGUATION

Word sense disambiguation (WSD) is the selection of the appropriate senses of a word in a given context. The need of natural language processing motivates the use of WSD in many crucial applications such as [4]:

a. **Machine translation:** appropriate translation of words depending on the context.

b. **Information retrieval and hypertext navigation:** retrieval of documents with occurrence of words in the most appropriate sense during keyword search.

c. **Content and thematic analysis:** generate appropriate category for a class of words that indicate a specific concept, idea, theme, on appearance through a text.

d. **Grammatical analysis:** speech tagging process for a given text.

Speech processing: generation of correct pronunciation of words during speech synthesis.

f. **Text processing:** spelling correction.

3. PROBLEM DEFINITION

Word sense disambiguation (WSD) involves the bonding of a given word in a text or discourse with a definition or meaning which is distinguishable from other meanings as well as potentially characteristic to that word. Words can have different senses. Some words have multiple meanings. This is called Polysemy. For example, the word 'bank' can be inferred in two different contexts:

a. A financial institution for managing money

b. A sloping part of earth besides the river [5][6]

Sometimes two completely different words are spelled the same. This is called Homonymy.

For example: Can, can be used as

a. Model verb: We **can** do this task.

b. Noun: A container: Open the **can** of juice. [7]

The imperfect distinction between polysemy and homonymy is handled by the word sense disambiguation (WSD) techniques. WSD determines the target sense a word in the present context, that has multiple of distinct senses in a given sentence. Though this process seems obvious for a human translator, development of algorithms to produce this human ability is an engineering challenge.

The task of determining sense for a target word necessarily involves two steps according to Ide and Veronis [4].

1. Determination of all the different senses for every word relevant to the text or dialogue under consideration, i.e., to select a sense inventory, e.g., from the lists of senses in dictionaries or from the synonyms in a thesaurus.
2. Assigning appropriate sense to each occurrence of a word in context. This involves matching the context of an occurrence of the target word with information from external knowledge sources or with contexts of previously disambiguated instances of the word.

Finally, an additional third step is to teach the computer to associate a word sense with a word in context using either machine learning or by manually created rules or metrics [7]. Few practical applications of WSD are given as :

a. a **search engine** may use it to determine what a user wants to search in its indices
b. a **program that converts speech to text** can use it to determine the exact spelling and pronunciation of a word
c. a **program making a concept map from a transcript** can use it to identify central topics more clearly.

In word sense disambiguation, electing the most frequent sense for an ambiguous word is a powerful heuristic. However, its usefulness is restricted by the availability of sense-explanatory data [9].

4. PROPOSED WORK

For evaluating the performance of search engines to handle sense ambiguity, a sufficiently large dataset is required. The precision and recall capacity of the search engine from this dataset against a query can be a performance measuring parameter. Each search engine retrieves and stores a massive collection of electronically stored texts and documents. Hence we adopt the individual database of each search engine as the prime source of corpora for ranking the senses of the words [8]. Upon querying a search engine, the user gets an ordered list of results ranked by a combination of factors. These results contain relevant results, less relevant results and results irrelevant to the queried keyword.

The higher is the number of relevant matches in the result list, the more precise is the searching algorithm. In this paper, all the ten keywords used, infer different senses (meanings) for different contexts.

For querying, we create a list of keywords having ambiguous senses from the Oxford dictionary database [10] in Table – 1. Each keyword in this list infers to different meanings or senses presented below. These keywords may also have other meanings additional to the specified ones. But in this paper, we consider only the most prominent senses for each keyword [10].

The relevance of the retrieved search results can be divided into following classes:

1. **Relevance according to sense-1** : if the result matches closely to sense-1 of the keyword
2. **Relevance according to sense-2** : if the result matches closely to sense-2 of the keyword
3. **Irrelevant or less relevant results** : if the result matches least or does not match at all to any of the two senses of the keyword

These keywords are consumed as search queries for retrieving results contained in Table – 2, 3 and 4. A match for every category is allotted a score of 1 and is placed in its appropriate category. The values of the first hundred results for these matches are collected in a table individual for each search engine. The table for each search engine contains following columns:

a. **Keyword:** The keyword of the ambiguous sense to be queried
b. **Total number of results retrieved:** Total results retrieved by the search engine. These results are fetched from the individual database of each search engine.
c. **Rank of most relevant result in search for sense-1:** The position on search results page at which the first sense of the word is detected for first time.
d. **Rank of most relevant result in search for sense-2:** The position on search results page at which the second sense of the word is detected for first time.
e. **Number of sense-1 results:** Results matching to first sense of the query keyword
f. **Number of sense-2 results:** Results matching to second sense of the query keyword
g. **Number of Irrelevant / Less relevant results:** Number of matches that least or does not match at all to any of the two senses of the keyword
h. **Precision:** Computed precision value of the search engine for the give query word
i. **F1 score:** Computed F1 score value of the search engine for the give query word.

The aggregate results for each search engine are later used to compute precision, recall and the F1 score.

Table – 1 : List of query keywords with ambiguous senses [10]

WORD	SENSE – 1	SENSE – 2
Bank	a financial establishment that uses money deposited by customers for investment, pays it out when required, makes loans at interest, and exchanges currency	the land alongside or sloping down to a river or lake
Land	Part of earth not covered by water	come down through the air and rest on the ground or another surface
Mind	element of a person that enables them to be aware of the world and their experiences, to think, and to feel; the faculty of consciousness and thought	be distressed, annoyed, or worried by
Mine	Digging, excavation to obtain something	Belonging to or associated to me (self)
Check	examine (something) in order to determine its accuracy, quality, or condition, or to detect the presence of something	a pattern of small squares
Band	a group of people who have a common interest or purpose	a flat, thin strip or loop of material, used as a fastener, for reinforcement, or as decoration
Form	printed document with blank spaces for information to be inserted	visible shape or configuration of something
Play	perform on (a musical instrument)	represent (a character) in a theatrical performance or a film
Right	moral or legal entitlement to have or do something	morally good, justified, or acceptable
File	a collection of information about a particular person or thing	a tool with a roughened surface or surfaces, typically of steel, used for smoothing or shaping a hard material.

We compute the F-score for each search engine, by using the retrieved mean precision value of each search engine. Precision is the ratio of number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage [1].

$$\textbf{Precision} = \frac{\textbf{Sum of the scores of sites retrieved by search}}{\textbf{Total number of sites selected for evaluation}}$$

The precision is a relative value computed for each search engine. It can be considered as the ration of the intersection of relevant and retrieved documents to the retrieved documents. Mathematically, the formula of computing the precision score can be expressed as follows :

$$\textbf{Precision} = \frac{|\{\textbf{Relevant Docs}\}| \cap |\{\textbf{Retrieved Docs}\}|}{|\{\textbf{Retrieved Docs}\}|}$$

Recall factor is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\textbf{Recall} = \frac{|\{\textbf{Relevant Docs}\}| \cap |\{\textbf{Retrieved Docs}\}|}{|\{\textbf{Relevant Docs}\}|}$$

Here we assume the recall factor to be 1. With an independent index database, every search engine is assumed to be recalling a complete set of relevant results from its indices, leading to a recall factor of 1. Further we compute the weighted harmonic mean of precision and recall, the F-measure, also known as the F1-score

$$\textbf{F1 score} = \frac{\textbf{2. Precision . Recall}}{\textbf{(Precision + Recall)}}$$

Upon processing the search results for the ten keywords,

Table – 2 : Values for First 100 Search Results of Google

Keyword	Total No. of Results Retrieved	Rank of Most Relevant Result in Search		Number of Sense-1 Results	Number of Sense-2 Results	Irrelevant / Less Relevant Results	Precision	F1-score
		Sense-1	Sense-2					
Bank	2,120,000,000	1	0	98	0	2	0.98	0.98
Land	2,820,000,000	2	15	73	1	24	0.74	0.85
Mind	2,010,000,000	1	55	67	1	32	0.68	0.8
Mine	1,120,000,000	7	16	67	5	28	0.72	0.83
Check	5,300,000,000	1	5	83	6	11	0.89	0.94
Band	1,950,000,000	1	22	90	5	5	0.95	0.97
Form	3,350,000,000	1	31	89	11	0	1	1
Play	6,730,000,000	9	15	72	8	20	0.8	0.88
Right	6,090,000,000	1	3	91	6	3	0.97	0.98
File	4,940,000,000	2	5	98	1	1	0.99	0.99
AVERAGE	**3,643,000,000**	**2.6**	**16.7**	**82.8**	**4.4**	**12.6**	**0.872**	**0.922**

Table – 3 : Values for First 100 Search Results of Bing

Keyword	Total No. of Results Retrieved	Rank of Most Relevant Result in Search		Number of Sense-1 Results	Number of Sense-2 Results	Irrelevant / Less Relevant Results	Precision	F1-score
		Sense-1	Sense-2					
Bank	02,68,000	1	0	97	0	3	0.97	0.98
Land	40,80,000	1	6	97	1	2	0.98	0.98
Mind	34,70,000	1	2	92	1	7	0.93	0.96
Mine	06,81,000	1	3	88	2	10	0.9	0.94
Check	21,70,000	2	4	92	3	5	0.95	0.97
Band	30,00,000	1	6	91	3	6	0.94	0.96
Form	16,30,000	1	3	95	4	1	0.99	0.99
Play	03,64,000	2	5	93	3	4	0.96	0.97
Right	78,30,000	1	3	86	10	4	0.96	0.97
File	04,10,000	1	3	99	1	0	1	1
AVERAGE	**23,90,300**	**1.2**	**3.5**	**93**	**2.8**	**4.2**	**0.958**	**0.972**

Table – 4 : Values for First 100 Search Results of Yahoo

Keyword	Total No. of Results Retrieved	Rank of Most Relevant Result in Search		Number of Sense-1 Results	Number of Sense-2 Results	Irrelevant / Less Relevant Results	Precision	F1-score
		Sense-1	Sense-2					
Bank	4,950,000	1	0	98	0	2	0.98	0.98
Land	5,110,000	1	3	98	1	1	0.99	0.99
Mind	4,230,000	1	2	94	1	5	0.95	0.97
Mine	0,613,000	1	4	89	3	8	0.92	0.95
Check	4,560,000	2	5	94	2	4	0.96	0.97
Band	2,820,000	1	3	91	4	5	0.95	0.97
Form	1,620,000	1	3	94	5	1	0.99	0.99
Play	0,420,000	2	3	93	3	4	0.96	0.97
Right	7,910,000	1	3	83	12	5	0.95	0.97
File	0,603,000	2	4	98	2	0	1	1
AVERAGE	**3,283,600**	**1.3**	**3**	**93.2**	**3.3**	**3.5**	**0.965**	**0.976**

retrieved from the 3 search engines, we have achieved the results presented above. These results are further analyzed and plotted as graphs in Figure – 1, 2, 3 and 4 for comparisons.

5. RESULTS AND OBSERVATIONS

This study reviewed in July 2016, evaluates precision and F-score accuracy of the search engines. Analyzing the above values, we arrive at following conclusions.

Precision of search engines has a heavy influence of the popularity of the keywords. Results reveal that Yahoo tops the precision ranking for ambiguous sensed words, while Google has the lowest precision of 0.872. Specifically, in case of Google, the density of irrelevant results towards lower part of search results was higher for every hundred results.

Acronyms, brand-names, trademarks increase complexity for the ranking mechanisms of the search engines. Misspells for any word uniformly affect the results of all search engines. For example, 'cheque' is misspelled to 'check', and displayed evenly by all 3 search engines in results.

Bing and Yahoo display a better dictionary approach for the results of the query words. Google displays maximum definition based results at prominent positions from wkipedia.org, as PageRank being its major ranking component heavily consumes the hyperlink structure of this website as input.

Irrelevant Results for each keyword have maximum effect on Google's output. The effect of inclusion of acronyms, brand names, trademarks in its indices is elevated by the heavy size of its database with more and more such results. As its ranking mechanism is now more inclined towards handling 'complex multi-word queries' and reformation of words, Google's single word query results are comparatively inferior to those of Bing and Yahoo. For example, on searching for 'mine', Google suggests to go for word 'mining', but displays fewer precise definitions exactly for 'mine' than Yahoo. Bing currently having the smallest database among the 3 services majorly contains dictionary definitions for each word and hence handles ambiguous words as the best of the three machines.

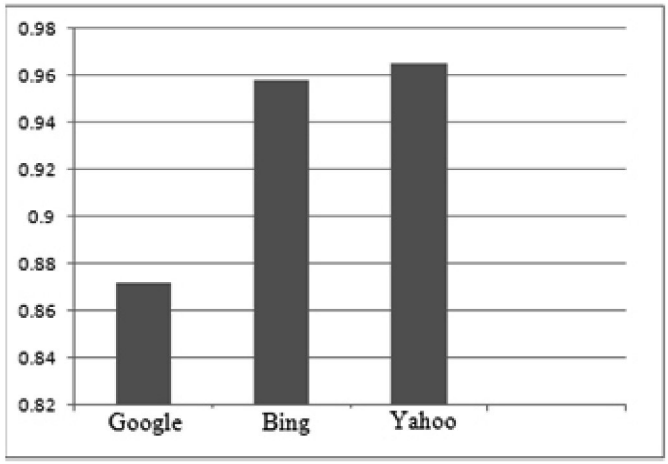

Figure-1 : Precision Comparison Graph

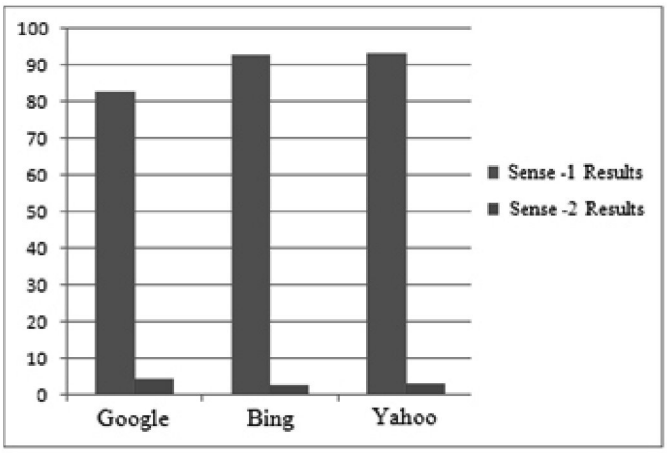

Figure-3 : Sense-1 / Sense-2 Relative Results Graph

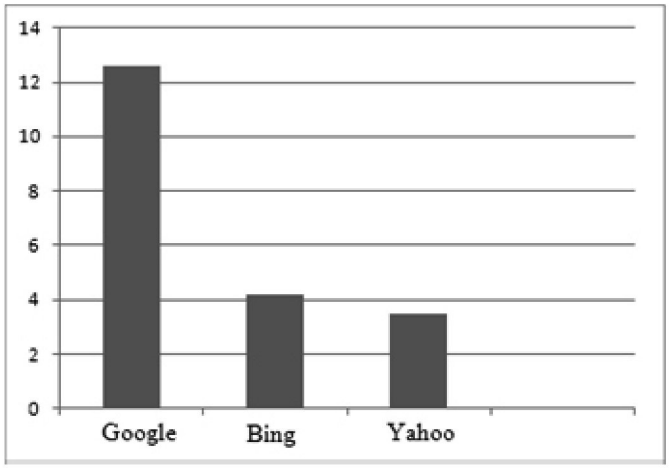

Figure-2 : Irrelevant Results Comparative Graph

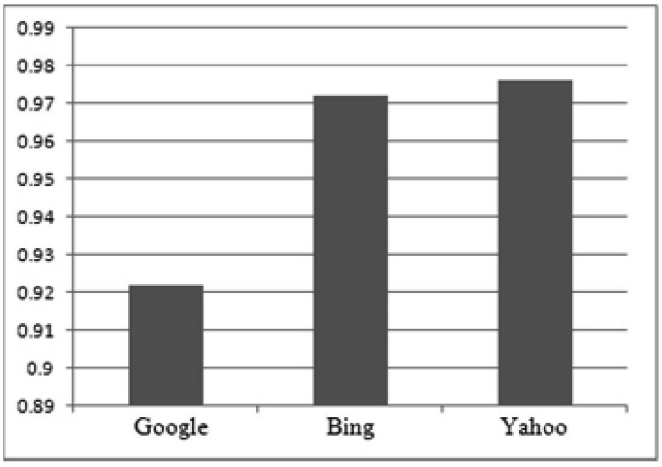

Figure-4 : F1 score comparision graph

7. CONCLUSIONS AND FUTURE WORK

Relative ranking of senses by each search engine is approximately common. Results for the dominant and more popular sense **o**f a keyword have attained more than 90% ranks in each set of hundred results. For example, the word 'form' has attained maximum results that lead to sense 'document'. Only Google, because of its vast database, showed a comparative improvement of 10% (approximate) in handling the inferior sense of 'appearance' for the word 'form'. Google displays a better management of paid inclusions without disturbing the design of their search results page and the relevancy of the content fetched compared to others.

F1 Score of the 3 search engines is a measure of their precision in handling ambiguous queries. Currently, despite of its moderate index size, Yahoo has topped the F1 score among these 3 services with a score of 0.976. Bing is second with 0.972, but an efficient competitor to Yahoo in F1 score precision test. With PageRank as its prominent ranking component and massive database, Google has ultimate potential to handle any query pattern with best relevance of results as depicted in figure-3, currently with an F1 score of 0.922. In future the accuracy of these results can be re-analyzed for a larger set of keywords with an automated script mechanism.

As evident from Table – 1, only single word queries are used as input in this paper. But the same process can be employed for multiword queries also. With the ever growing volume of data with all three search engines, a better filtration mechanism of aged and less relevant websites, acronyms, brand names, trademarks in its indices is also required for higher precision among results. A hybrid approach of human classification with automated ranking can uplift the quality of search results for a search machine while keeping pace with accuracy and speed of computations.

REFERENCES

[1] Beitzel.,Steven M. **On Understanding and Classifying Web Queries,** Submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy, Illinois Institute of Technology; May 2006.

[2] Usmani Tauqeer A., Pant D., Bhatt Ashutosh K.; **A Comparative Study of Google and Bing Search Engines in Context of Precision and Relative Recall Parameter**; International Journal on Computer Science and Engineering (IJCSE); Vol. 4 No., January 2012

[3] Brin S., Page L.; **The Anatomy of a Large-Scale Hypertextual Web Search Engine**; 7th International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

[4] Ide N., Véronis J.; **Word sense disambiguation: The state of the art**; Computational Linguistics; 1998;

[5] Navigli R.; **Word Sense Disambiguation: A Survey**; ACM Computing Surveys, Vol. 41, No. 2, Article 10, February 2009.

[6] Craggs, David J.; **An Analysis and Comparison of Predominant Word Sense Disambiguation Algorithms**; Thesis Edition; Faculty of Computing, Health and Science, Edith Cowan University; 2011

[7] Palta Esha; **Word Sense Disambiguation: First Stage Report**; Submitted in partial fulfilment of the requirements of the degree of Master of Technology; Indian Institute of Technology, Bombay; May 2007;

[8] Mihalcea R., Moldovan Dan I.; **A Method for Word Sense Disambiguation of Unrestricted Text**; ACL '99 Proceedings of the 37th annual meeting of the Association of Computational Linguistics. pages 152–158; 1999;

[9] Lapata M., Keller F.; **An Information Retrieval Approach to Sense Ranking**; Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pages 348–355; 2007;

[10] Oxford Dictionaries Online; http://oxforddictionaries.com

◆ ◆ ◆