Impact Of Word Sense Ambiguity For English Language In Web IR

Prachi Gupta[1], Dr.AnuragAwasthi[2], RiteshRastogi[3]
[1,2,3]*Department of computer Science and engineering,*
Noida institute of engineering and technology,Greater noida

Abstract— **Word Sense ambiguity for English language is considered as the major restraint in any natural language processing applications especially in web IR. Determining the word sense ambiguity is one of the most important problem in Web IR. Basically an information retrieval system is a system which retrieves only relevant documents according to user needs. In this paper, we analyse the impact of word sense ambiguity for English Language as it plays an important role and have great impact on search engine performance as users are not aware of the ambiguity problem when they are searching on any search engines.**

***Keywords*: - Precision, Recall, Word Sense Disambiguation, Search Engines.**

1. INTRODUCTION

An Information retrieval system is a system that returns as many relevant documents as possible. Basically, an IR system is finding documents that match the information need description. There are many different methods for evaluating the performance of web IR. The method basically requires a collection of documents and query and every document is known to be either relevant or non-relevant to a particular query.

The method given below is used for the evaluation of search engine:

PRECISION (P): It 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" [1]. Chu & Rosenthal [2] used a three-level relevance score (relevant, somewhat relevant, and irrelevant) while Gordon and Pathak [3] used a four-level relevance judgment (highly relevant, somewhat relevant, somewhat irrelevant, and highly irrelevant).

1.1 WORD SENSE AMBIGUITY IN WEB IR

An IR system is impacted by the characteristics of text, one such characteristic is word sense ambiguity. Most words are ambiguous in nature, what sense a word occurrence has depends on the context it appears in. For some words, their senses are unrelated, for example like the word 'bat' could refer to an implement used in sports to hit balls or a flying mouse like animal. For most words however, their senses are related (e.g. through metaphor), the word 'crash' for example can refer to a physical event or the value of shares in a stock market dropping. A number of users had tried to retrieve articles about the Prime Minister using the query 'major'. This query caused many articles about 'John Major' to be retrieved, but in addition many more articles were retrieved where 'major' was used as an adjective or as the name of a military rank. Somehow, when an ambiguous word is used in a sentence, people are usually able to select the correct sense of that word without conscious effort. The manual Word Sense Disambiguating (WSD) ability has been investigated, an overview of which can be found in Hirst[4]. Choueka and Lusignan[5], working with the French language, found that people could accurately determine the sense of a particular word from reading the previous two words alone. Miller [6]briefly describes similar work by Kaplan using the English language which seems to draw similar results to those of Choueka and Lusignan. These works show that accurate disambiguation can be performed without exposure to the wider context in which an ambiguous word appears. Many words have several meanings or senses. For such words there is ambiguity about how they should be interpreted. Word Sense Disambiguation (WSD) is the task of examining word tokens in context and specifying exactly which sense of each word is being used. The main problem of word sense disambiguation is deciding what the senses are for a particular word.

2. RELATED WORK

Search Engines are the basic tools of Information retrieval on the web. There are two problems in using words to represent document contents and query in information retrieval: ambiguity and different words which represent the same concept. These problems can be addressed by using query

expansion. They focused on analysing the implementation of query expansion, word sense disambiguation (WSD), iterated relevance feedback, and some retrieval variations to retrieval performance [7].

Various researchers have studied the effect of ambiguity problem on the performance of information retrieval task on English queries. According to Sanderson in 1994 showed short queries are mostly benefited from the ambiguity resolution [8]. His work showed disambiguation lead to better performance. Lesk in 1986 proposed the algorithm for WSD; he also implemented his algorithm on the short text sample and found the good results [9].

Krovetz and Croft [10] in 1992 studied the relationship between sense mismatch and irrelevant documents. They concluded that the co-occurrence of multiple words interacting within a query naturally performs some element of disambiguation indicating that disambiguation might only be of benefit over short queries. The experiments show that there is considerable ambiguity even in a specialized database. Word senses provide a significant separation between relevant and non relevant documents, but several factors contribute to determining whether disambiguation will make an improvement in performance.

To assess the role of automated word sense disambiguation to improve retrieval effectiveness work has been done by C.M Stokoe and Pr J.Tait. They found only small increase in R-.precision [11]. R.Song and Z.Luo in an effort to define and differentiate ambiguous query a supervised learning approach has been proposed to automatically identify ambiguous queries. Main idea was to report a document with a vector of semantic categories by applying the query ambiguity clarifier. It was estimated that 16% sampled queries are ambiguous in a real query log.

Sanderson [12] used artificial pseudo-words [13] to attempt to measure the effects of ambiguity on the Cranfield and TREC-B collections. By introducing ambiguous terms into these collections he measured the retrieval performance and evaluated the results against the baseline for the original collection. He found that queries consisting of “one or two terms” were heavily affected by ambiguity.

Sanderson [14] says that word sense ambiguity only recently became regarded as a problem to information retrieval which was potentially solvable. The growth of interest in word senses resulted from new directions taken in disambiguation research. Although the majority of attempts to improve retrieval effectiveness were unsuccessful, much was learnt from the research. Most notably a notion of under what circumstance disambiguation may prove of use to retrieval.

Sanderson [14] returned to the problem of WSD and IR in 2000 when he offered three key factors that affect WSD for IR. Firstly, skewed distribution of senses and collocation query effects are the reason why ambiguity has only a small impact on IR performance. Secondly, in order to benefit from automated WSD you need highly accurate disambiguation. This statement is less precise than his 1994 conclusions. Automatic word sense disambiguation has long been studied: by Gale, Church and Yarowskywork dating back to 1950. For many years, disambiguators could only accurately disambiguate text in limited domains or over a small vocabulary. In recent years, however, the situation changed with large improvements in scalability resulting in the possibility of applying a disambiguator to accurately resolve the senses of words in a large heterogeneous corpus.

Finally, he concludes that simple dictionary or thesaurus based word representations have not been shown to offer improvements in IR and as such he advocates the use of broader semantic groupings.

Schütze and Pederson [15] remains one of the clearest indications to date of the potential for WSD to improve the precision of an IR system. Their technique involved examining the context of every term in the TREC 1 category B collection and clustering them based entirely on the commonality of neighbouring words. The idea behind this is that words used in a similar sense will share similar neighbours, and by building a vector spaced representation of this co-occurrence and identifying different directions in the model we can indicate different contexts.

3. IMPACT OF WORD SENSE AMBIGUITY ON SEARCH ENGINES

Word Sense Ambiguity is an open problem of natural language processing that governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when there are multiple meanings of a word (polysemy). Word sense ambiguity is not something that we encounter in every day life, except in the context of jokes.

Ambiguity in natural language has long been recognized as having a detrimental effect on the performance of text based information retrieval (IR) systems. Sometimes called the polysemy problem [16], the idea that a word form may have more than one meaning is entirely discounted in most traditional IR strategies. If only documents containing the relevant sense of a word in relation to a particular query were retrieved this would undoubtedly improve precision.

Ambiguity is rarely a problem for humans in their day to day communication, except in extreme cases. Most words have many possible different meanings. A computer program has no basis for knowing which one is appropriate or not, even if it is obvious to a human.

Word sense ambiguity is a topic that has been studied for many years in the Information Retrieval (IR) community, starting with Weiss’s small scale experiments [17] through to a more thorough examination of the topic in the 1990s. Most

of the past disambiguation research focused on ambiguity of words found in dictionaries, which have poor coverage proper nouns or phrases such as titles and names etc. This is unfortunate since it is increasingly clear that names of people, locations, organizations, acronyms, etc, are common queries in search engines. Some of these nouns will have high levels of ambiguity, but the extent of the ambiguity is little understood. Word sense ambiguity in natural language has long been recognized as having a detrimental effect on the performance of text based information retrieval (IR) systems. Sometimes called the polysemy problem, the idea that a word form may have more than one meaning is entirely discounted in most traditional strategies.

The ambiguity in natural language is considered as the major barrier in language processing applications, especially in information retrieval. Some query terms have a clear cut sense in their query. However some query terms hold ambiguity. Identifying the appropriate sense of the words in the given context is a difficult job for the search engines. Word sense disambiguation gives solution to the many natural language processing systems including information retrieval.

Sense ambiguity in queries is clearly understood by an example:

Query: Today is **cold.**

In this the word “cold” has two senses -

Sense 1 = Disease

Sense 2 = temperature

Therefore, in this query the “cold” is a polysemous word.

Following are the lists some polysemous words with their different senses or meanings:

TABLE 1
LIST OF AMBIGUOUS WORDS WITH THEIR SENSES

WORDS	SENSES
Cold	Disease, temperature
Sign	Visible clue, zodiac sign
Case	Term used in court, portable container for carrying objects
Interest	Related in terms of money or interest in any work
Figure	Diagrams, digits in math
Close	Come together, end

Sense ambiguity is one of the major problems in Information Retrieval on web. Many words are polysemous in nature. Identifying the appropriate sense of the words in the given context is a difficult job for the search engines. Word sense disambiguation gives solution to the many natural language processing systems including information retrieval.

We took 30 TREC queries which are ambiguous in nature and have shown the effect of ambiguity on the performance of the search engines. Following are the Example of ambiguous queries.

Examples:

1. Wall **paint** is blue.
2. The train is standing on the **platform**.
3. **Forestry** is a field of study.
4. There are four **seasons** in a year.
5. Build a **bat** house

The above queries are examined on the search engine the result is shown below in the Table 2.

TABLE 2

PRECISION OF GOOGLE IN CONTEXT OF SENSE AMBIGUITY PROBLEM FOR ENGLISH LANGUAGE

Query	Doc. Retreived	Precision@20
1	140,000,000	0.44
2	31,600,000	0.66
3	2,860,000	0.37
4	175,000,000	0.55
5	2,550,000	0.5
6	1,020,000,000	0.55
7	18,400,000	0.66
8	435,000,000	0.33
9	2,210,000	0.75
10	662,000,000	0.37
11	4,420,000	0.22
12	325,000	0.44
13	12,600,000	0.62
14	9,260,000,000	0.44

15	16,200,000	0.5
16	338,000,000	0.55
17	174,000,000	0.66
18	335,000,000	0.55
19	45,100,000	0.44
20	683,000,000	0.75
21	374,000,000	0.33
22	187,000,000	0.77
23	3,150,000	0.44
24	374,000,000	0.33
25	95,000,000	0.44
26	363,000,000	0.55
27	66,000,000	0.37
28	78,998,000	0.75
29	123,000,000	0.44
30	112,342,000	0.87

The precision of the queries after removing ambiguous words of the ambiguous queries i.e. we remove the ambiguous words and make the queries unambiguous and calculate the precision @20. The unambiguous words queries are examined on the search engine the result is shown below in Table 3.

TABLE 3
PRECISION OF GOOGLE AFTER REMOVING AMBIGUOUS SENSES FOR ENGLISH LANGUAGE

Query	Doc. Retreived	Precision@20
1	140,000,000	0.54
2	31,600,000	0.78
3	2,860,000	0.58
4	175,000,000	0.62
5	2,550,000	0.62
6	1,020,000,000	0.58
7	18,400,000	0.68
8	435,000,000	0.38
9	2,210,000	0.78
10	662,000,000	0.42
11	4,420,000	0.48
12	325,000	0.46
13	12,600,000	0.69
14	9,260,000,000	0.52
15	16,200,000	0.52
16	338,000,000	0.58
17	174,000,000	0.64
18	335,000,000	0.58
19	45,100,000	0.6
20	683,000,000	0.78
21	374,000,000	0.58
22	187,000,000	0.8
23	3,150,000	0.56
24	374,000,000	0.38
25	95,000,000	0.46
26	363,000,000	0.58
27	66,000,000	0.40
28	78,998,000	0.78
29	123,000,000	0.5
30	112,342,000	0.9

5. RESULTS

The fig. 1 shows the graph of both ambiguous queries and unambiguous queries of precision in Table 2 and Table 3.The graph below shows that the precision is low when the query is ambiguous and the precision is high when queries are unambiguous i.e. sense ambiguity also affects on the performance of search engines.

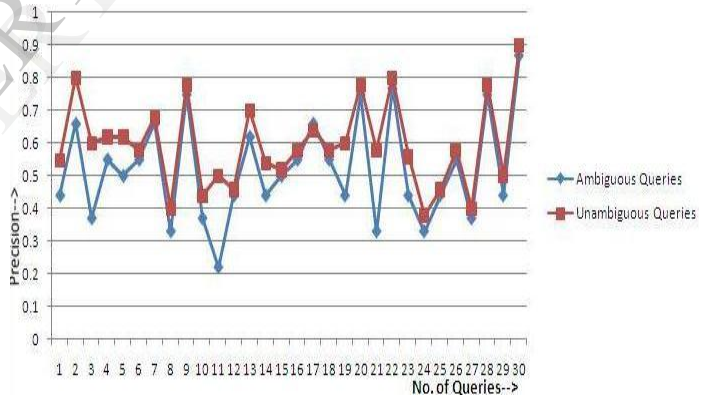

Fig.1 . Average Precision comparison between ambiguous and unambiguous queries References

The queries used in Table 2 are ambiguous as per Word Net senses [18]. We have replaced the ambiguous words of each of these queries to make them unambiguous and find out the precision in Table 3.The precision in table 3 when the queries are unambiguous are high as compare to the precision in table 2 when the queries are ambiguous. From this evaluation it is clear that the search engine performance is greatly affected by the sense ambiguity.

6. CONCLUSION

When we calculate the precision of both ambiguous queries and unambiguous queries the precision of unambiguous queries are high that means we are getting more relevant results. For this we take 30 ambiguous queries and find out the precision@20after that we removed ambiguous words and make them unambiguous then also we calculate precision@20. From both graph it is clear that the precision of unambiguous words queries are high. So, it is clear that word sense ambiguity affected the search engine performance and search engine itself is not capable to cope up this problem.

7. FUTURE SCOPE

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 it is require to know the impact of word sense ambiguity on the performance of search engine .These results are used for any word sense detection algorithm which in turn used for word sense disambiguation algorithm because without prior detection automatic disambiguation may lead to the wastage of computational power.

REFERENCES

[1] 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.

[2] Chu, H., & Rosenthal, M. (1996). Search engines for the World Wide Web: a comparative study and evaluation methodology. In Proceedings of the 59th annual meeting of the American Society for Information Science (pp. 127-135). Medford, NJ: Information Today.

[3] Gordon M, Pathak P, Finding information on World Wide Web: the retrieval effectiveness of search engines, Information Processing and Management 141-180, 35(1999)

[4] C. Stan fill & B. Kahle (1986). Parallel free text search on the connection machine system., in Communications of the ACM, 29(12): 1229-1239.

[5] Y. Choueka & S. Lusignan (1985). Disambiguation by short contexts, in Computers and the Humanities, 19: 147-157.

[6] G. A. Miller (1954). Communication, in Annual Review of Psychology, 5: 401-420.

[7] Paskalis, F.B.D.(2011), "Word sense disambiguation in information retrieval using query expansion" In Proceedings of the International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, pp 1-6.

[8] 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.

[9] Lesk, M; (1986); Automatic sense disambiguation using machine readable dictionaries". Proceedings of the SIGDOC, Toronto, ON, Canada, pp. 24-26..

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

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

[12] Sanderson, M. "Word Sense Disambiguation and Information Retrieval" In Proceedings of the 17th International ACM SIGIR, Pp 49 – 57, Dublin, IE, 1994.

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

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

[15] Schütze, H; Pederson, J. O. "Information Retrieval Based on Word Senses" In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, Pp 161 – 175, Las Vegas, NV, 1995.

[16] Kowalski, G; Maybury, M. "Information Storage and Retrieval Systems Theory and Implementation" Kluwer, Pp 97, 2000.

[17] Allan, J., Carterette, B., Aslam, J., Pavlu, V., Dachev, B., Kanoulas, E. (2007) Million Query Track 2007 Overview, in TREC 2007 Notebook.

[18] http://muse.dillfrog.com/ambiguous_words.php