Journal of Basic and Applied Engineering Research
Print ISSN: 2350-0077; Online ISSN: 2350-0255; Volume 1, Number 8; October, 2014 pp. 1-5

http://www.krishisanskriti.org/jbaer.html

A Survey on Impact of Word Sense Ambuiguity on Search Engine's Performance

Arti Mishra[1], Meenakshi Pathak[2]

[1,2]*Computer science Department, SRMSWCET, Bareilly*

Abstract: **Word sense ambiguity is rarely a problem for humans in their day to day communication, except in extreme cases. Ambiguity is an open problem of natural language processing, which governs the process of identifying which sense of word i.e. meaning is used in a sentence, when the word has multiple meanings (polysemy). Most words have many possible meanings, for example: a word "cold" refer to a disease and it also refers to "temperature". A computer program has no basis for knowing which one is appropriate, even if it is obvious to a human. This paper covers the comprehensive analysis of web queries in English language to know the impact of ambiguity on the various search engines like-Google search engine, Yahoo search engine, etc. Our result shows that the performance of search engines is quiet affected by the sense ambiguity problem. We disambiguate the web queries by dividing it in two parts: ambiguous and unambiguous web queries, and show their affect on the performance of the search engine.**

Keywords: **Word sense ambiguity, search engine, natural language, web queries, natural language processing.**

1. INTRODUCTION

Word sense ambiguity means a single word or sentence is interpreted differently by different users. The main reason for this is that a single word has more than one meaning (exact meaning depends on the context). 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 polysemy problem [1], 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. 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 been long 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.

As an example, consider the following two sentences:

1. I can hear ***bass*** sounds.
2. Today is ***cold***.

Any system that tries to find out the meanings of the two sentences will need to represent somehow different senses for the word bass and cold. In the first sentence, the word bass refer to a type of fish and it also refers to a tone. In the second sentence the word cold refer to temperature and also a disease. This is ambiguity in the queries with respect to the senses and it also affects the results of the user.

1.1 Types of Ambiguity:

(i). Lexical ambiguity occurs when a word has several meanings. For example, bank.

(ii). Syntactic ambiguity, also called *structural ambiguity*, occurs when a given sequence of words can be given more than one grammatical structure, and each has a different meaning. In the terminology of compiler construction, syntactic ambiguity occurs when a sentence has more than one parse. For example

1. The Tibetan history teacher can be read as The (Tibetan history) teacher or The Tibetan (history teacher). [Analytical ambiguity]
2. The police shot the rioters with guns. [Attachment ambiguity]
3. I saw Peter and Paul and Mary saw me. [Coordination ambiguity]
4. Perot knows a richer man than Trump. [Elliptical ambiguity]

(iii). Semantic ambiguity occurs when a sentence has more than one way of reading it within its context although it contains no lexical or structural ambiguity. For example, all linguists prefer a theory.

(iv). Pragmatic ambiguity occurs when a sentence has several meanings in the context in which it is uttered. For example, every student thinks she is a genius.

2. LITERATURE REVIEW

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 [2] through 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, names, etc. This is unfortunate since it is increasingly clear that names of people, locations, organizations, acronyms, etc. are common queries in search engines.

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 [3]. His work showed disambiguation lead to better performance.

Sanderson [4] used artificial pseudo-words [5] 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.

3. EVALUATION METHODOLOGY

The sense 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.

The table 1 given below contains the list of some polysemous words with their different senses or meanings:

Table 1. List of Polysemous Words with Their Senses

WORDS	SENSES
Paint	Painting as art, wall coverings
Platform	Base, Political platform
Forestry	Forestry service, field of study
Seasons	Weather, performance
Bat	Mammal, Baseball bat
Cold	Disease, temperature
Sign	Visible clue, zodiac sign
Case	Term used in court, portable container for carrying objects
Mouse	Device, rat, cartoon
Interest	Related in terms of money, interest in any work
Figure	Diagrams, digit in math
Close	Come together, end
Right	Law, correct, direction on right side
Score	Marks, grade, written form of musical composition
Balance	Remaining money, state of equilibrium
Break	Interval time, separation, breaking of tissue
Bank	River, financial institution
Pound	Nontechnical unit of force, unit of money
Exercise	A task performed or problem solved, activity of exerting your muscles
Function	Party, or math term
Pitch	Range of voice, cricket pitch
Present	Period of time, birthday present
Dry	Spray dry, or hanging up wet
Bug	Error, mistake, Insect

Sense ambiguity is one of the measure problems in Information Retrieval on web. Many words are polysemous in nature. We took 30 TREC queries which having ambiguous words and these are ambiguous queries and in place of ambiguous word we place related sense word, so these queries are unambiguous in nature and have shown the effect of ambiguity on the performance of the search engines.

Table2. Set of ambiguous and unambiguous queries

Query No.	Ambiguous queries	Unambiguous queries
1	Wall paint is blue.	Wall color is blue.
2	The train is standing on the platform.	The train is standing on the railway platform.
3	Forestry is a field of study.	Forestry service is a field of study.
4	There are four seasons in a year.	There are four cycles in a year.
5	Build a bat house.	Build a bat mammal house.
6	Today is cold.	Today is cold temperature.
7	There are 12 sign in	There are 12 zodiac sign

	astrology.	in astrology.
8	This case is very critical.	This situation is very critical.
9	Clip of light bulb.	Clip art of light bulb.
10	Bank of India.	State Bank of India.
11	A bug terminates a program.	An error terminates a program.
12	Python are found mostly in rainy season.	Python snakes are found mostly in rainy season.
13	Mouse is favorite cartoon of kids.	Mickey Mouse is favorite cartoon of kids.
14	I have an interest in science.	I like to study science.
15	Draw the figure of a flower.	Draw the diagram of a flower.
16	Close the door.	Shut the door.
17	There should be a break between two lectures.	There should be a gap between two lectures.
18	Please turn right.	Please turn right direction.
19	Score of team India in World cup.	Runs of team India in World cup.
20	There is no balance in my phone.	There is no remaining money in my phone.
21	The river is dry.	The river is empty.
22	Always live in present.	Always live in today.
23	My aim is to become a doctor.	My dream is to become a doctor.
24	Pound is money.	Pound is a unit of money.
25	The pitch of sound is high.	The cricket pitch of sound is high.
26	Use of cosine function.	Use of cosine math function.
27	Exercise is necessary to keep your body fit.	Physical Exercise is necessary to keep your body fit.
28	The chair of ACL conference is Prof. S.K.D.	The chair person of ACL conference is Prof. S.K.D.
29	It is a major accident.	It is a big accident.
30	Law of motion	Law of motion

In the above English queries given in Table 2, the results is calculated by retrieving total documents and calculate the precision on top 10 documents (precision@10), by using precision method.

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" [6]. Chu & Rosenthal (1996) [7] used a three-level relevance score (relevant, somewhat relevant, and irrelevant) while Gordon and Pathak (1999) [8] used a four-level relevance judgment (highly relevant, somewhat relevant, somewhat irrelevant, and highly irrelevant).

The precision can be calculated by the formula shown below:
Precision = tp / (tp+fp)
Where,
tp – true positive result, means that the retrieved documents are true or relevant according to the user query.
fp – false positive result, means that the retrieved documents are unexpected or irrelevant according to the user query.
The above queries are examined on the search engine the result is shown below in Table2.

Table 2. PRECISION OF GOOGLE IN CONTEXT OF SENSE AMBIGUITY PROBLEM FOR ENGLISH LANGUAGE

Query	Doc. Retrieved	Precision of ambiguous queries(@10)	Precision of unambiguous queries(@10)
1	140, 000, 000	0.44	0.49
2	31, 600, 000	0.66	0.68
3	2, 860, 000	0.37	0.39
4	175, 000, 000	0.55	0.59
5	2, 550, 000	0.5	0.58
6	1, 020, 000, 000	0.55	0.56
7	18, 400, 000	0.66	0.68
8	435, 000, 000	0.33	0.44
9	2, 210, 000	0.75	0.78
10	662, 000, 000	0.37	0.4
11	4, 420, 000	0.22	0.32
12	325, 000	0.44	0.54
13	12, 600, 000	0.62	0.67
14	9, 260, 000, 000	0.44	0.48
15	16, 200, 000	0.5	0.58
16	338, 000, 000	0.55	0.59

17	174, 000, 000	0.66	0.68
18	335, 000, 000	0.55	0.62
19	45, 100, 000	0.44	0.48
20	683, 000, 000	0.75	0.77
21	374, 000, 000	0.33	0.44
22	187, 000, 000	0.77	0.77
23	3, 150, 000	0.44	0.44
24	374, 000, 000	0.33	0.44
25	95, 000, 000	0.44	0.55
26	363, 000, 000	0.55	0.66
27	66, 000, 000	0.37	0.44
28	78, 998, 000	0.75	0.77
29	123, 000, 000	0.44	0.55
30	112, 342, 000	0.87	0.9

4. DISCUSSION

We have done an extensive analysis of the impact of ambiguity issues of web queries. The Google though has been capable of searching very efficiently still not very capable of understanding user's intension and the context of queries. A minor change in the sense of polysemous word term (at least from the searchers point of view) may result in considerable change in precision.

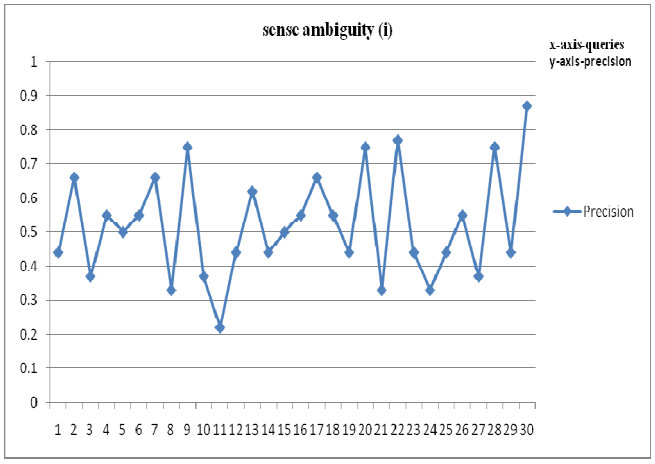

Fig1. Performance of search engine (Google) with ambiguous queries

The fig.1 shows the graph of ambiguous queries given in Table 2 and shows that the precision is low when the query is ambiguous. Therefore, sense ambiguity also affects on the performance of search engines (like Google). The search engine is not capable to cope up this problem.

The 30 queries that are used in Table2 above have an ambiguous word as per Word Net sense [25].We have replaced these ambiguous words of each of these queries to make them unambiguous queries. The above fig.2 shows the graph of unambiguous queries and shows that the precision is high when the queries are unambiguous. From this evaluation it is clear that the search engine performance is greatly affected by the sense ambiguity.

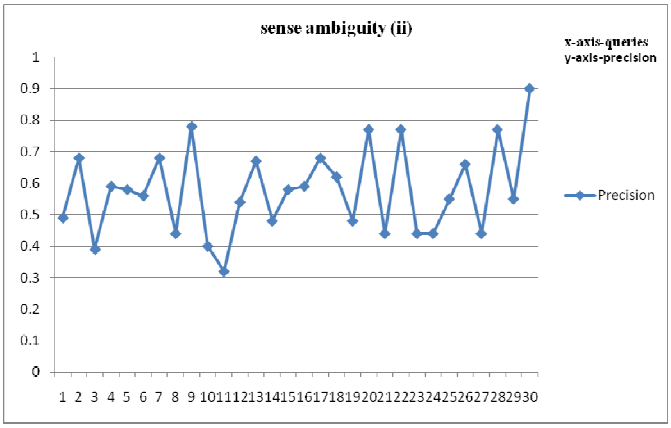

Fig2. Performance of search engine (Google) with unambiguous queries

5. CONCLUSION

The issues discussed in this paper towards the ambiguity query formation at the end user level are generally ignored by common web searchers. Our result conclude that the performance of the search engines is quiet affected by the sense ambiguity problems. Ambiguity is 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 this paper, we made an effort to highlight in the results show that the performances of the search engines are affected word sense ambiguity. The query term ambiguity may sometimes drastically reduce the relevancy of a search engine. The ambiguity detection and disambiguation of web queries are essential which affect on the performance of search engine.

REFERENCES

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

[2] Allan, J; Carterette, B; Aslam, J; Pavlu, V; Dachev, B; Kanoulas, E.

[3] 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.

[4] Sanderson, M., "Word sense Disambiguation and Information Retrieval", Proceedings of 17th International ACM SIGIR, Pp 49-57, Dublin, IE, 1994.

[5] Yarowsky, D. "One Sense Per Collection" In Proceedings of ARPA Human Language Technology Workshop, Pp 266-271, Princeton, NJ, 1993.

[6] 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.

[7] 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.

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