ABSTRACT

The ambiguity in word senses has been recognized as a major challenge for the information retrieval systems. Hindi language web information retrieval, like other languages, faces the problem of sense ambiguity. The sense ambiguity problem deteriorates the performance of every natural language processing (NLP) application. The performance of Hindi language web information retrieval is also affected by it. In this paper, the author formalized an approach for the disambiguation of the senses to improve the performance of Hindi web information retrieval. Our system works in such a way that ambiguity detection has been performed before disambiguation of web queries. Test samples of 100 queries have been selected. When these queries were subjected to ambiguity detection, we found that 43% of them have been detected unambiguous. After ambiguity detection, the disambiguation approach is followed which is based on HSC (Highest Sense Count). Query disambiguation approach further follows query expansion. The expanded query generates the new result set which results into high precision and high similarity score. The 57 expanded queries are tested against 1000 test document instances. The overall improvement is 45% in the average precision, 23% in interpolated average precision and a significant improvement in the average similarity score of the new generated result set. The overall accuracy of our approach has been 61.4% and it improves the performance of the system by 45%.

Keywords: Highest Sense Count (HSC), Hindi Language, Query Expansion, Web Information Retrieval, Word Sense Disambiguation

INTRODUCTION

The Internet is the foremost source of information for the human population. Though, English is the most dominated and preferred language for the web access globally, the rapid growth in the popularity of Internet in non-English speaking countries like India (the Internet penetration in India is one amongst the fastest in the world), have increasingly made the need and importance of reaching out to the non-English speaking zone. Being the national language of India, a majority of people use Hindi as a first language. With the increase in contents written in native languages on the Internet, a proper mechanism is needed to make this content noticeable and
available wherever and whenever necessary. The Hindi language web IR face numerous challenges like – unavailability of standard keyboard, less amount of contents, encoding standards and, of course, excellence search engines. Besides these, another challenge for search engines is sense ambiguity problem. The problem of ambiguity sometimes also called as polysemy problem (Kowalski & Maybury, 2000) and it refers that that a word form may have more than one meaning and therefore, in an IR setup the relevancy of retrieval will certainly improve if only documents containing the relevant sense of a (ambiguous) word are retrieved for a particular query.

Ambiguity in natural languages has long been recognized as having an adverse effect on the performance of information retrieval (IR) systems in general and web IR in particular (Stokoe, Oakes & Tait, 2003).

The motivation behind this work is to implement WSD in Hindi language web IR to improve its performance. In order to achieve this task of disambiguation of web queries in Hindi, the structure of study has been as follows,

For a given set of web queries in Hindi, first step that the system performs is to detect the queries for ambiguity. So, only the queries detected ambiguous have to be passed to WSD algorithm for disambiguation. In the second step, these ambiguous queries have been disambiguated using the Highest Sense Count (HSC) based approach. Finally, based on disambiguation result of the sense, the original query has been expanded. The query expansion usually appends the correct sense of the ambiguous word in the original query. After this final step, the expanded query is resubmitted to the search engine (Google has been used in this research) for performance comparison of retrieved results.

The research contributions in WSD and related work has been discussed in the second section of this paper, and the next section discusses sense ambiguity in web IR and subsequent sections discuss the detailed approach used in this research.

**RELATED WORK**

Some of the early researches in WSD and its integration with IR can be found in the works of Krovetz and Croft (1992), Sanderson (1994), Sanderson (2000), Gonzalo, Peñas, and Verdejo (1999). These contributors justified the significance of WSD in the area of IR. They just broke the myth of the earlier researchers like Zernik (1991), Voorhees (1993), Wallis (1993), Sussna (1997), who concluded their work by illustrating that there is no significance of WSD on the improvement of performance of IR systems.

The majority of work done in Hindi language is restricted to the Machine Translation. Key researchers like Bhattacharyya (Bhattacharyya, Sinha, Kumar, Pande & Kashyap, 2004) who proposed the statistical approach which was very near to Lesk (1986) approach. Another unsupervised approach was given by Neetu Mishra (Mishra, Yadav & Siddiqui, 2009) for Hindi language WSD. In another work, Klapaftis and Manandhar (Klapaftis & Manandhar, 2005) used the Total Sense Score (TSS) for the disambiguation.

Besides that, some other researchers also have used web documents for the disambiguation approach (Gaona, Gelbukh & Bandyopadhyay, 2009; Katsiouli & Kalamboukis, 2009). The similar kind of work is performed by Navigili and Crisafulli (2010) to improve the web search results. Their work faces the problem of higher computational cost. The most time-consuming phase of their approach is the construction of the query graph, which requires intensive querying of the database of co-occurrences calculated from the Web1T corpus.

Wok in web query expansion has also been reported by many, such as Gong, Cheang and Hou (2005). They used WordNet to combine the query expansions along each semantic dimension as their overall solution. However, Using only WordNet or the web as whole for the query expansion may not be a feasible solution for the query expansion in WSD. Since sense disambiguation is dependent on the context of the terms therefore it is quiet justified that the context of the query terms can be identified.