Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 2, Issue. 5, May 2013, pg.139 – 144

RESEARCH ARTICLE

GENETIC ALGORITM BASED HINDI WORD SENSE DISAMBIGAUTION

Sabnam Kumari[1], Prof. (Dr.) Paramjit Singh[2]
[1]Department of Computer Science and Engineering, PDM College of Engineering, Bahadurgarh, Haryana, India
[2]Professor of Computer Sciences, PDM College of Engineering, Bahadurgarh, Haryana, India

[1] Shabnam022@gmail.com; [2] director_engg@pdm.ac.in

Abstract— Word Sense Disambiguation (WSD) is a problem of computationally determining which "sense" of a word is activated by the use of the word in particular context. To figure out the appropriate meanings of polysemous nouns in the given context Genetic Algorithm is used. This is crucial for various applications like 'machine translation', 'speech processes' and 'information retrieval' etc. while the work on WSD for English is voluminous, to our knowledge, and this is the first attempt of using Genetic Algorithm for Hindi language. Wordnet for Hindi developed at IIT Bombay, a lexical knowledge base for Hindi, is used. The main focus is on removing the ambiguity of the sense using the context by applying Genetic Algorithm.

Key Terms: - Ambiguity; Genetic Algorithm; Hindi Wordnet; Sense Disambiguation; Seed; Word Sense Disambiguation

I. INTRODUCTION

Word Sense Disambiguation (WSD) is the task of finding the appropriate sense of a word used in a given sentence, when the word may have more than one sense. E.g.

- **मंहगाई से हर वर्ग के लोग परेशान हैं।** Here **वर्ग** is interpreted as 'class'.
- **सात का वर्ग उनचास होता है।** Here **वर्ग** is interpreted as 'square of number'.
- **यह पाँच सैन्टीमीटर का वर्ग है।** Here **वर्ग** is interpreted as 'square shaped figure'

Some words may not be easy to disambiguate as they may have multiple senses that are close to each other. In some senses disambiguation may be impossible altogether using the given sentence. Now consider **यह तो वर्ग है।**

Given only the above sentence, one may translate it as " This is a square shaped figure" or "This is the square of the number" or "This is a class". All are having valid senses of the word **वर्ग**.

In another example **सोना सोना चाहती है।**

This sentence can be interpreted as "Sona wants gold" or "Sona wants to sleep" or "Gold wants to sleep" or "Sleep wants Gold" etc. Thus these scenarios require a look at context of the discourse to disambiguate between the possible senses.

WSD is a task of classification: word senses are the classes, context provides the evidence, and each occurrence is assigned to one or more of its possible classes based on evidence [1]. Sense Disambiguation [2] is an 'intermediate task' which is not an end itself, but rather is necessary at one level or another to accomplish most NLP tasks. Sense Disambiguation involves Sense Knowledge. Sense Knowledge can be represented by a vector, called a sense knowledge vector *(sense ID, features)*, where features can be either symbolic or empirical. The word to be sense tagged always appears in a context. Context can be represented by a vector, called a context vector (*word*, *features*). Thus, we can disambiguate word sense by matching a sense knowledge vector and a context vector.

II. APPROACHES TO WSD

As in all natural language processing, there are two main approaches to WSD – deep approaches and shallow approaches [1].

[A] Deep Approaches

Deep approaches presume access to a comprehensive body of world knowledge. E.g. consider the word "bass" with two distinct senses: 'a type of fish' and 'tones of low frequency'. Knowledge such as "you can go fishing for a type of fish, but not for low frequency sounds" and "songs have low frequency sounds as parts, but not types of fish" is used to determine in which sense the word is used. These approaches are not very successful in practice, mainly because we don't have access to such a body of knowledge, except in very limited domains. However, if such knowledge did exist, then deep approaches would be much more accurate than the shallow approaches [1][6].

There are two types of Deep approach of Word Sense Disambiguation are:

- Selectional restriction- based approaches
- Approaches based on general reasoning with 'world knowledge'

[B] Shallow Approaches

Shallow approaches don't try to understand the text. They just consider the surrounding words, using information like "if 'bass' has words 'sea' or 'fishing' nearby, it probably is in the fish sense; if 'bass' has the words 'music' or 'song' nearby, it is probably in the music sense." These rules can be automatically derived by the computer, using a training corpus of words tagged with their word senses. This approach, while theoretically not as powerful as deep approaches, gives superior results in practice, due to our limited world knowledge. The different types of Shallow approaches of WSD are:

- Dictionary-based approaches.
- Machine learning approaches
- Supervised methods
- Semi-supervised
- Unsupervised methods
- Hybrid approach

III. HINDI WORDNET [3]

Wordnet is a freely available semantic lexicon for the English and Hindi language whose design is inspired by current psycholinguistic theories of human lexical memory. Wordnet for Hindi is produced by people researching in the Centre for Indian Language Technology (CFLIT), IIT-B, under the direction of Prof. Pushpak Bhattacharya. Its design is inspired by the famous English Wordnet. In Wordnet, each part of speech word (nouns/verbs...) is organized into taxonomies where each node is a set of synonyms called synsets. Each synset represents a specific meaning. It includes the word, its definition (gloss), its explanation, and its synonyms.

Each Entry in Hindi Wordnet consists of {synsets, gloss, and ontology}. It defines a number of semantic relations for Hindi word based on its possible senses. It also defines the following semantic relations to connect synsets.

- **Hypernym**: Y is hypernym of X if every X IS-A (KIND-OF) Y.
- **Hyponym**: Y is hyponym of X if every Y IS-A (KIND-OF) X.
- **Meronym**: Y is a meronym of X if Y is a part of X.
- **Holonym**: Y is a holonym of X if X is a part of Y.
- **Antonym**: Y is antonym of X if X is opposite of Y.
- **Attribute**: Y is attribute of X is Y is a value of X.

For example, The synset {**पेड़, वृक्ष, पादप, तरू, विटप**} has the hypernym relation to {**जड़, मूल, सोर**}, a hyponym relation to {**कदंब**}, the meronym relation to {**शाखा**}, a holonym relation to {**जंगल**}, an attribute relation to {**फलदार**}. The synset {**मोटा**} has an antonym relation to {**पतला**} .

Current Status of Hindi Wordnet is still under construction. In the version 1.0 is an attempt to cover all the common concepts in Hindi. The present status is as follows:

Total unique words: 93584

Total Synsets: 37391

Linked Synsets: 24319

IV. RELATED WORK

Manish Sinha, Mahesh Kumar Reddy .R, Pushpak Bhattacharyya , Prabhakar Pandey and Laxmi Kashyap [4], worked on "Hindi Word Sense Disambiguation" that was the first attempt for an Indian language at automatic WSD. The approach is to compare the context of the word in a sentence with the contexts constructed from the Wordnet and chooses the winner. The output consisted of a particular synset number designating the sense of the word. The evaluation was done on the Hindi corpora provided by the Central Institute of Indian Languages.

Neetu Mishra, Shashi Yadav and Tanveer J. Siddiqui [5], "An Unsupervised Approach to Hindi Word Sense Disambiguation" developed an Algorithm that learns a decision list using untagged instances.

Some seed instances are provided manually. Stemming has been applied and stop words have been removed from the context. The list is then used for annotating an ambiguous word with its correct sense in a given context. The evaluation has been made on 20 ambiguous words with multiple senses as defined in Hindi Wordnet.

Rohan Sharma [6], "Word Sense Disambiguation for Hindi Language" made an attempt to resolve the ambiguity by making the comparisons between the different senses of the word in the sentence with the words present in the synset form of the Wordnet and the information related to these words in the form of parts-of-speech.

Parul Rastogi and Dr. S.K. Dwivedi [7], "Performance comparison of Word Sense Disambiguation (WSD) Algorithm on Hindi Language Supporting Search Engines" compared the performance of WSD Algorithm by using Highest Sense Count (HSC). The Hindi language search engines face the problem of sense ambiguity. The objective is comparative analysis of the WSD algorithm results on the three Hindi language search engines- Google, Raftaar and Guruji.

Neetu Mishra and Tanveer J. Siddiqui [8], "An Investigation to Semi-Supervised approach for Hindi WSD", investigated Yarrowsky algorithm. After elimination of both, stemming and stop words, the maximum observed precision is 61.7 on 605 test instances.

Sandeep Kumar Vishwakarma and Chanchal Kumar Vishwakarma [9], "A Graph Based approach to Word Sense Disambiguation for Hindi Language" combined Lesk semantic similarity measures and Indegree algorithms for graph centrality and 65.17% accuracy has been obtained.

V. METHODOLOGY

Genetic Algorithm (GA) is a heuristic search algorithm used to find approximate solutions to optimization and search problems using techniques inspired by evolutionary biology. This observation was first mathematically formulated by John Holland in 1975 in his paper, "Adaptation in Natural and Artificial Systems". According to Koza, "the fact that the genetic algorithm operates on a population of individuals, rather than a single point in the search space of the problem, is an essential aspect of the algorithm. The population serves as the reservoir of the probably-valuable genetic material that the crossover operation needs to create new individuals with probably-valuable new combinations of characteristics". GA simulates the natural evolution mimicking processes the nature uses – selection, crossover, mutation and evaluation. It is based on the Darwin's principle 'Survival of the Fittest' [10]. In nature, competition among individuals for scanty resources results in the fittest individuals dominating over the weaker ones. In GA, a population of chromosomes (abstract

representations of candidate solutions to an optimization problem) evolves toward better set of solutions. An additional advantage of the genetic algorithm is that the problem solving strategy involves using the strings fitness to direct the search; therefore they do not require any problem-specific knowledge of the search space, and they can operate well on search spaces that have gaps, jumps, or noise.

VI. PROPOSED WORK

The proposed work consists of developing a method used to resolve semantic ambiguity for Hindi language by using Genetic Algorithm. The algorithm proceeds incrementally. The steps are summarized below:

- Formulate initial population
- Initialize population
- Evaluate fitness function
- Perform selection
- Reproduction
- Crossover
- Mutation
- Generate new population

Base for evaluating fitness function is WU-Palmer similarity to find relatedness of the words.

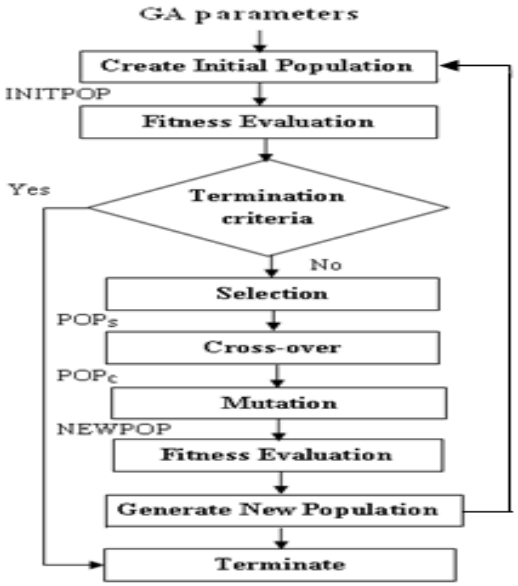

Fig 1: Work Flow of Genetic Algorithm

First we need to decide and set the GA parameters like chromosome length, population size and cross-over probability etc. Then we generate the initial population. Generally it is a random population. After the initial population is generated, we calculate the fitness value of each chromosome. Then we repeat the steps- Selection->Crossover->Mutation->Evaluation until termination criteria is satisfied.

VII. CONCLUSION AND FUTURE WORK

In this paper we haved the Hindi Wordnet for fundamental task, viz. disambigaution of Hindi words. To our knowledge, by using Genetic Algotithm, no attempt has been made in the past to adrees the problem of word sense disambigaution on Hindi language. The algorithm i.e. Genetic Algorithm gives the optimized results after disambigaution. Till now we have some intermediary results which are under process. Our system will currently deal with the nouns only. In future, words of other parts of speech can be included. With the ennrichement of the algorithm the system performance is expected to be very impressive.

ACKNOWLEDGEMENT

Prof. (Dr.) Paramjit Singh is the professor in Department of Computer Science and Engineering at PDM College of Engineering, Bahadurgarh, Haryana. I am especially grateful for his guidance and contributions by generously giving his time and carefully reviewing this manuscript.

REFERENCES

[1] Edmonds Philip Glenny, Eneko. Agirre, Word sense disambiguation: algorithms and applications, Springer, 2006, online, Available: books.google.co.in/books?isbn=1402048092.
[2] Reddy Siva, Inumella, Abhilash, Rajeev Sangal, Soma Paul, "All Words Unsupervised Semantic Category Labeling for Hindi" Proceedings of the International Conference RANLP, Borovets, Bulgaria, pages 365–369, September 2009.
[3] Pande P. Bhattacharya P, S. Jha, D. Narayan. A Wordnet for hindi,2001.
[4] Sinha Manish, Reddy Mahesh Kumar, Bhattacharyya R Pushpak, Pandey Prabhakar and Kashyap Laxmi, "Hindi Word Sense Disambiguation", Indian Institute of Technology Bombay, Department of Computer Science and Engineering Mumbai, 2008.
[5] Mishra Neetu, Yadav Shashi and Siddiqui Tanveer J., "An Unsupervised Approach to Hindi Word Sense Disambiguation," Indian Institute of Information Technology, Allahabad. UP, India, 2009.
[6] Sharma Rohan, "Word Sense Disambiguation For Hindi language" Thapar University Patiyala, CSE Dept., India, 2007.
[7] Rastogi Parul and Dr. S.K. Dwivedi, "Performance comparison of Word Sense Disambiguation (WSD) Algorithm on Hindi Language Supporting Search Engines", International Journal of Computer Science Issues, vol. 8, issue.2, March 2011.
[8] Mishra Neetu and Siddiqui Tanveer J., "An Investigation to Semi-Supervised approach for Hindi Word sense Disambigaution", Proceedings of Trends in Innovative Computing 2012- Intellligent System Design, 2012.
[9] Vishwakarma Sandeep Kumar and Vishwakarma Chanchal Kumar, "A Graph Based approach to Word Sense Disambiguation for Hindi Language", International Journal of Scientific Research Engineering & Technology (IJSRET), vol1, issue 5, pp 313-318, Aug. 2012.
[10] Holland, J. H. (1975), Adaptation in Natural and Artificial Systems. University of Michigan Press. Ann Arbor, MA.
[11] WordNet terminology (WordNet Glossary) in WordNet 1.2 browser
[12] "Word Sense Disambiguation", 2009, http://en.wikipedia.org/wiki/Wordsense_disambiguation# Approaches_and_methods

