Hindi Word Sense Disambiguation

Manish Sinha Mahesh Kumar Reddy .R Pushpak Bhattacharyya
Prabhakar Pandey Laxmi Kashyap

Department of Computer Science and Engineering
Indian Institute of Technology Bombay, Mumbai
India
{manish, mahesh, pb,pandey,yupu}@cse.iitb.ac.in

Abstract

Word Sense Disambiguation (WSD) is defined as the task of finding the correct sense of a word in a specific context. This is crucial for applications like Machine Translation and Information Extraction. While the work on automatic WSD for English is voluminous, to our knowledge, this is the first attempt for an Indian language at automatic WSD. We make use of the Wordnet for Hindi developed at IIT Bombay, which is a highly important lexical knowledge base for Hindi. The main idea is to compare the context of the word in a sentence with the contexts constructed from the Wordnet and chooses the winner. The output of the system is a particular synset number designating the sense of the word. The mentioned Wordnet contexts are built from the semantic relations and glosses, using the Application Programming Interface created around the lexical data. The evaluation has been done on the Hindi corpora provided by the Central Institute of Indian Languages and the results are encouraging. Currently the system disambiguates nouns. Work is on for other parts of speech too.

Keywords

Hindi Wordnet, Text Similarity Measures, Semantic Relations in the Wordnet, Intersection Similarity

1. Introduction

Word Sense Disambiguation (WSD) is defined as the task of finding *the correct* sense of the word in a context. The task needs large amounts of word and word knowledge. Let us consider the word सम्बन्ध in the following Hindi sentence:

> ऋग्वेद की एक ऋचा में दस्यु के विशेषणों से उनकी संस्कृति एवं वेदिक समाज के साथ उनके **सम्बन्ध** पर पूर्ण प्रकाश पड़ता हे। उन्हें अक्रतु, मृध्रवाच, अश्रद्ध, पणि, अयज्ञ आदि कहा गया हे।

Figure 1.1: One of the possible usage of सम्बन्ध

From the Hindi Wordnet*, we find that there are 6 senses of सम्बन्ध, *viz,*

1. संबंध, **सम्बन्ध**, मतलब, नाता, ताल्लुक, वास्ता, रिश्ता - *किसी प्रकार का लगाव या संपर्क:"इस काम से राम का कोई संबंध नहीं है"*
2. संबंध कारक, षष्ठी, संबंध, **सम्बन्ध**, सम्बन्ध कारक - *व्याकरण में वह कारक जिससे एक शब्द का दूसरे शब्द के साथ संबंध सूचित होता है:"संबंध कारक की विभक्ति का, के, की, रा, रे री आदि हैं जैसे यह किस की पुस्तक है?"*
3. लगाव , संबंध, **सम्बन्ध**, संसर्ग - *दो वस्तुओं में किसी प्रकार का लगाव या संपर्क बतलाने वाला तत्व:"साथ रहते-रहते तो जानवरों से भी लगाव हो जाता है"*
4. संबंध, **सम्बन्ध**, रिश्ता - *विवाह अथवा उसका निश्चय:"मंगला के लिए बिलासपुर में संबंध पक्का हो गया है"*
5. संबंध, **सम्बन्ध** - *एक साथ बँधने, जुड़ने या मिलने की क्रिया:"प्रेम-भाव से आपसी संबंधों में प्रगाढ़ता आती है"*
6. नाता, रिश्ता, संबंध, **सम्बन्ध** - *मनुष्यों का वह पारस्परिक संबंध जो एक ही कुल में जन्म लेने अथवा विवाह आदि करने से होता है:"मधुरिमा से आपका क्या नाता है?"*

Figure 1.2: Senses of सम्बन्ध obtained from the Wordnet

In this particular case, sense 1 is the most appropriate one, though sense 5 and 6 too are relevant.

1.1 Related Work for English

Yarowsky proposed a solution to WSD using the thesaurus and a supervised learning approach [3]. Word

* Hindi Wordnet [8] is an important lexical resource developed at IIT Bombay, India.

associations are recorded and for an unseen text, the senses of the words are detected from the learnt associations. Aggire and Rigau uses a measure based on the proximity of the text words in Wordnet (Conceptual Density) to disambiguate the words [4]. The idea that translation presupposes WSD is given by Nancy lde. to disambiguate words using bilingual corpora [2]. The design of the well-known work-bench for sense disambiguation for *WASP* is described by Kilgarriff [5]. Lin [16] and Lesk [17] have studied theoretical definitions of similarity and provided word similarity measures- which are hypernymy based and gloss based respectively.

2. Wordnet Principle

Wordnet is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory [1]. Each word meaning can be represented by a set of word-forms known as *synonym set*s or *synset*s. Synsets are created for content words, *i.e.*, for Noun, Verb, Adjective and Adverb.

2.1 Lexical Matrix

The following table- called *Lexical Matrix*- is an abstract representation of the organization of lexical information. Word–forms are imagined to be listed as headings for the columns and word meanings as headings for the rows. Rows express *synonymy* while columns express *polysemy*.

Word Meanings	Word-Forms F_1 F_2 F_3 ….. F_n
M_1	$E_{1,1}$ $E_{1,2}$
M_2	$E_{2,2}$
M_3	$E_{3,3}$
……	…..
M_m	$E_{m,n}$

Table 2.1: Illustrating the concept of Lexical Matrix

For example, the synset {**कलम**, **पेन**, **क़लम**, **लेखनी**} gives the meaning **उपकरण जिसकी सहायता से कागज़ आदि पर लिखते हैं**. **कलम** belongs to a synset whose members form a row in the lexical matrix, and the row number gives a unique id to the synset. **कलम** has another meaning- **पेड़ की वह टहनी जो दूसरी जगह बैठाने या दूसरे पेड़ में पैबंद लगाने के लिए काटी जाए**- which comes in the column headed by this word.

2.2 Semantic Relations in Wordnet

Hindi Wordnet design is inspired by the famous English Wordnet [1]. The basic semantic relations are as follows:

Relation	**Meaning**
Hypernymy/Hyponymy	Is-A (Kind-Of)
Entailment/Troponymy	Manner-Of (for verbs)
Meronymy/Holonymy	Has-A (Part-Whole)

Table 2.2: Illustrating the nature of the relations in Wordnet

For instance, we have the synset {**घर, गृह**}. The hypernymy relation (Is-A) of it links to {**आवास, निवास**}. Its meronymy relation (Has-A) links to {**आँगन**} {**बरामदा**} and {**अध्ययन कक्ष**} and hyponymy relation to {**बाड़ी**}, {**सराय**} and {**झोपड़ी**}.

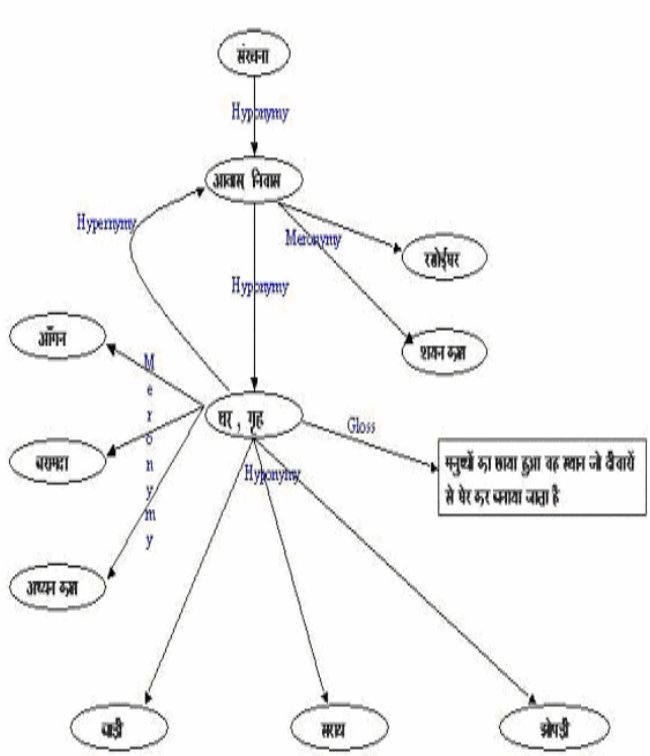

Figure 2.1: A small part of the Hindi Wordnet

2.3 Wordnet Application Programming Interface

The WSD task needs various information from the Wordnet, which in turn calls for the availability of an Application Programming Interface to the Wordnet. Figure 3.1 shows the organization of the API. To take a particular example, *findtheinfo()* function receives input arguments as *word form*, *syntactic category*, *search type (e.g., hypernymy)* and *sense number*. This will return the *search type (i.e., hypernymy)* output in a buffered form.

These APIs are meant to do followings: *(1) Morphological Processing (2) Database Searching (3) Utilities.* Morphological processing routines extract the

stem from the word. Database search functions are used retrieve information from the Wordnet. Utilities are useful in other operations which might be useful to process words.

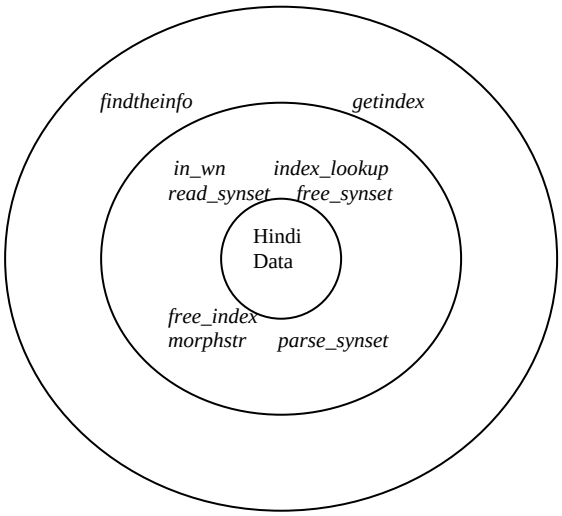

Figure 2.2: Layers of Application Programming Interface around the Wordnet

3. Methodology: Our Approach to WSD

We describe a statistical technique for assigning senses to words in Hindi. A word is assigned a sense with the use of (i) the context in which it has been mentioned (ii) the information in the Hindi Wordnet and (iii) the overlap between these two pieces of information. The sense with the maximum overlap is the *winner sense.*

WSD Algorithm: Finding the word's Correct Sense

1. For a polysemous word *w* needing diambiguation, a set of context words in its surrounding *window* is collected. Let this collection be *C*, *the context bag*.
2. For each sense *s* of *w*, do the following
   (a) Let *B* be the bag of words obtained from the
      (I) Synonyms
      (II) Glosses
      (III) Example Sentences
      (IV) Hypernyms
      (V) Glosses of Hypernyms
      (VI) Example Sentences of Hypernyms
      (VII) Hyponyms
      (VIII) Glosses of Hypernyms
      (IX) Example Sentences of Hypernyms
      (X) Meronyms
      (XI) Glosses of Meronyms
      (XII) Example Sentences of Meronyms
   (b) Measure the *overlap* between C and B using the intersection similarity measure.
3. Output that the sense *s* as the most probable sense which has the *maximum overlap*.

Figure 3.1 gives the pictorial description of the basic idea of the strategy. The idea behind using the intersection similarity measure is to capture the belief that there will be high overlap between the words in the context and the *related words* found from the wordnet lexical and semantic relations and glosses.

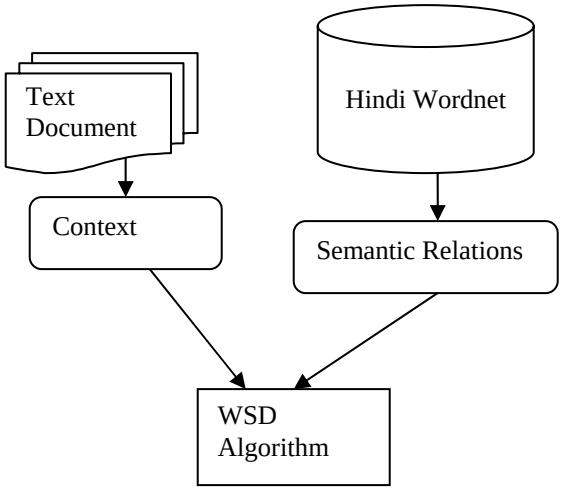

Figure 3.1 Extracting semantic relations from Wordnet and building context from the text for WSD

4. Components in the System

4.1 Parameters

- *Wordnet relations*: We have used *hypernymy*, *hyponymy* and *meronymy* relations. Since, these relations are semantic in nature; we obtain the synsets, their glosses and example sentences. We call the collection of words from words from Wordnet as the *Semantic Bag*.

- *Word Context Size*: The current sentence in which *w* is forms the most important context. We add to this the *previous* and the *following* sentences too. We call the collection of context words as the *Context Bag*.

4.2 Implementation Modules

- **BuildContext:** This module builds the context bag from the input document.

- **NounSemanticExtractor:** This module builds the semantic bag by exploiting the semantic relations in the Wordnet. Input to this module is the polysemous word.
- **Tokenizer:** This module finds the unique tokens from the input document. This is an intermediate module required by *BuildContext* and *NounSemanticExtractor*.
- **Intersection:** This computes the overlap between the two input bags.
- **Rank**: This ranks the senses according the *amount of intersection*.

The details of the functions used for the WSD task is given in Appendix-I.

5. Evaluation

We use the Hindi corpora from the Central Institute of Indian Languages (CIIL), Mysore as the test bed for sense disambiguation. We do this task currently for *nouns* only.

5.1 Test Document

Following is the part of a test document. The domain is sociology.

आर्य बाहर से आये या इसी देश के मूल निवासी रहे हैं इस पर पर्याप्त सामग्री पक्ष विपक्ष में प्रस्तुत की जाती है। लेकिन इस मूल प्रश्न को ही भुला दिया जाता है कि आर्य नाम की कोई जाति या प्रजाति भी रही है या नहीं। आर्य का जब ऋग्वेद और उसके परवर्ती साहित्य में कहीं प्रजाति अर्थ में प्रयोग है ही नहीं तो यह विवाद कहाँ से शुरू होता है इस पर हमने इसी शोध पत्रिका के विगत अंक में विचार किया है। वस्तुतः आर्य शब्द गुण परक है और श्रेष्ठ अर्थ में इसका प्रयोग होता है। श्रेष्ठता व्यक्तिगत चरित्र के साथ ही व्यवस्था दर्शन आचरण सम्पत्ति सिद्धान्त आदि के व्यापक सन्दर्भ में होने से इसके साथ प्रजाति विश्लेषण का भ्रम उस समय होना स्वाभाविक है जब हम साम्राज्यवादियों की इस मनोवृत्ति के माध्यम बन जाते हैं कि भारतवर्ष में एक राष्ट्र होने की क्षमता ही नहीं है। सभी वर्णों को उत्पन्न करने की क्षमता वाले इस देश के सम्बन्ध में भूगर्भशास्त्रियों का विशष्ठ वर्ग भी इस पक्ष में विचार प्रस्तुत कर रहा है कि भारतवर्ष में ही आदि सृष्टि हुई और यहीं से विश्व में मानस विकास हुआ।ऋग्वेद से लेकर मानवधर्मशास्त्र तक प्रजाति विश्लेषण के जो तथ्य मिलते हैं वे इसी के समर्थन में निर्विवाद रूप से हैं।वैदिक साहित्य में जो विचार व्यवहार और सिद्धान्तगत विभिन्नताएँ आती हैं और उनमें संघर्ष की भी स्थिति चलती रहती है उसपर आर्य अनार्य प्रजाति भेद करने की अपेक्षा हमें वैदिक अवेदिक भेद करना अधिक समीचीन प्रतीत होता है। वर्ण शब्द का प्रयोग रंग के अर्थ में ऋग्वेद में पर्याप्त प्राप्त होता है किन्तु उसका प्रयोग सामाजिक समूह के अर्थ में भी हुआ है। कृष्ण और शुक्ल वर्ण के परस्पर संघर्ष में यह भी ध्यान देना आवश्यक है कि ऐसा संघर्ष प्रजातिगत नहीं है जिसे आर्य द्रविड़ प्रजाति के रूप में विभक्त करते हुए उत्तर दक्षिण के संघर्ष का रूप दिया जाय जों देने की चेष्टा की गयी है। इनका सम्बन्ध आचरणगत भी है। एक ही परिवार में शुक्ल और कृष्ण वर्ण के भी होने के उल्लेख हैं। इसी आधार पर यह तर्क प्रस्तुत करना कि द्रविड़ों के देवता शिव और आर्यों के देवता विष्णु में समन्वय की प्रक्रिया रही है हास्यास्पद लगता है क्योंकि विष्णु कृष्ण वर्ण के हैं और शिव कर्पूर वर्ण के। इसमें तो विष्णु को शुक्ल और शिव को कृष्ण होना चाहिए था। यहाँ देवमंडल का विभाजन करने वाले वर्ण के स्थान पर शिव और विष्णु के व्यावहारिक स्वरूप को आधार बना देते हैं। इसी व्यावहारिक आधार पर यदि पूरा विश्लेषण देखा जाय तो आर्य प्रजाति न होकर आचरणगत श्रेष्ठता का प्रतीक होगा।ऋग्वेद में प्राप्त द्धरणों से स्पष्ट है कि वर्ण शब्द आचरण एवं व्यवस्तागत श्रेष्ठता से विशेष रूप में जुड़ा हुआ है और उसका प्रयोग जाति के अर्थ में भी हुआ है। एक स्थान पर ब्राह्मण को दिव्य और शूद्र को असुर कहा गया है। यहाँ असुर शब्द के प्रयोग में व्यापक रूप से गुण और सामाजिक समूह दोनों अर्थों का प्रयोग पाया जाता है। ऋग्वेद की ऋचा को तेत्तिरीय ब्राह्मण के साथ जोड़ कर देखने पर स्पष्ठ होता है कि असुर शब्द का प्रयोग जाति के अर्थ में हुआ है जो शूद्र के आचरण को ध्यान में रखकर किया गया है। इससे यह निष्कर्ष निकालने का खतरनाक प्रयास हुआ है कि शूद्र के लिए असुर शब्द के प्रयोग से स्पष्ट है कि यह जाति आर्यों अनार्यों के दासत्व के रूप में प्रविष्ट हुई है। वस्तुतः असुर शब्द का प्रयोग वरूण और इन्द्र के लिए भी हुआ है। असुर शब्द का कोई रूढ़ अर्थ लेना ही भ्रमात्मक है।दास दस्यु और आर्य भेद का तो प्रश्न ही नहीं है। कहींकहीं उल्लेख आता है जहाँ वेदिकों और दस्युओं में भेद करने के लिए इन्द्र से प्रार्थना की गयी है। यदि कृष्ण वर्ण के आधार पर दस्यु भिन्न होते तो इस प्रकार की प्रार्थना क्यों की जाती इतना अवश्य है कि दस्युओं के विरोध से वेदिकों की रक्षा के लिए बारबार प्रार्थना की गयी है। इन उद्धरणों से स्पष्ट होता है कि दस्युओं में शारीरिक नहीं आचारगत भिन्नता रही है। वेदिकों को बर्हिष्मत और दस्युओं को अयज्ञ कहा गया है। दस्यु और दास कहीं भिन्न और कहीं एक ही माने गये हैं। एक ही ऋचा में दोनों के नाम आये हैं और भिन्न अर्थ का द्योतन करते हैं। दोनों के समूह अवेदिक आचरण के रूप में पाये जाते हैं। वेदिकों के विपरीत संघर्ष में दोनों मित्र भी हो जाते हैं। फलतः वेदिको के लिए दोनों समान रूप से शत्रु हो गये और उनके विनाश के लिए समान रूप से इन्द्र से प्रार्थना की गयी है। दोनों के लिए आये विशेषणों से स्पष्ट है कि उनके वेदिक विरोध का कारण वेचारिक सेद्धान्तिक और व्यवस्थागत है। दोनों को अक्रतु कहा गया है जिसका तात्पर्य है कि दोनों उस विचारधारा के रहे हैं जो यज्ञ धर्म का विरोध करते हैं। ऋग्वेद की एक ऋचा में दस्यु के विशेषणों से उनकी संस्कृति एवं वेदिक समाज के साथ उनके सम्बन्ध पर पूर्ण प्रकाश पड़ता है। उन्हें अक्रतु मृध्रवाच अश्रद्ध पणि अयज्ञ आदि कहा गया है। इसमें पणि शब्द धन संचय के लिए आया है अर्थात् जो वेदिक विधि के अनुसार धन का वितरण नहीं करता है उसे भी दस्यु कहा गया है। अनेक स्थानों पर दासों को पणि कहा गया है। यहाँ जो भी विशेषण सामने आये हैं उनसे अत्यन्त स्पष्ट है कि सारा संघर्ष व्यवस्थागत है इसमें वर्ण या जाति का महत्वनहीं होता।उक्त विशेषणों से स्पष्ट है कि वेदिक व्यवस्ता के मूलाधार यज्ञीय जीवन का दस्यु और दास विरोध करते हैं और स्वंय वे उसे स्वीकार नहीं करते। वे वेदिक भाषा के अलावा अन्य भाषा का प्रयोग करते हैं। वे वेदिक व्यवस्था में श्रद्धावान् नहीं हैं। इन्हीं आधारों पर इन्हें अवेदिक कहा गया है। इन्द्र इन पर विजय प्राप्त करते हैं। वेदिक काल से ही राज्य का उद्देश्य वेदिक व्यवस्था की रक्षा करना रहा है। बारबार इन्द्र की प्रार्थना की गयी है कि दासों से आर्य व्यवस्ता की रक्षा के लिए वह युद्ध करे। मूल संघर्ष आर्य व्यवस्था का सर्वांगीण संरक्षण रहा है। दासों के साथ इस प्रकार के संघर्ष के विपुल प्रमाण पाये

जाते हैं।दासों के साथ वेदिकों के संघर्ष में अनेक कारणों में सम्पत्ति मुख्य हे। दासों को धनिनः कहा गया हे और इसी सन्दर्भ में पणि शब्द का भी प्रयोग होता हे। दास सम्पत्ति का संचय करते हैं।

The results obtained from this particular document are shown in table 5.1.

Word	Synset	Comment
विधि	ढंग, रीति, तरीक़ा, शैली, रीत, ढर्रा, विधि, पद्धति, तरीका, तौर, अंदाज़, अंदाज, कार्यविधि, क़ायदा, कार्य शैली	Correct
गुण	सद्गुण, अच्छाई, गुण, ख़ूबी, खूबी	Partially correct
राज्य	प्रदेश, राज्य, प्रांत, प्रान्त, सूबा	Incorrrect
पक्ष	दल, पार्टी, पक्ष	Correct
सम्बन्ध	संबंध, सम्बन्ध, मतलब, नाता, ताल्लुक़, ताल्लुक, वास्ता, रिश्ता, सरोकार	Correct
वर्ण	अक्षर, वर्ण, आखर, हरफ़, हर्फ़	Incorrect
अंक	अंक, नाट्यांक, नाटक अंक	Incorrect
भेद	अंतर, असमानता, फर्क़, भिन्नता, भेद, व्यतिरेक	Correct
अंश	खंड,अंश,टुकड़ा,भाग,हिस्सा,अंग, विभाग,कला,चरण	Correct
विद्या	कला, फ़न, हुनर, विद्या	Partially Correct
चेष्टा	प्रयत्न, प्रयास, उद्यम, उद्योग, यत्न, कोशिश, चेष्टा, पैरवी	Correct
नाम	नाम	Correct
विचार	विचार, ख़याल, मंतव्य	Partially Correct
रंग	रंग	Correct
धन	मूलधन, पूँजी, असल, मूल, धन	Partially Correct
विरोध	प्रतिवाद, खंडन, विरोध	Partially Correct
प्रयोग	प्रयोग	Incorrect
ध्यान	स्मृति, याद, सुध, सुधि, ध्यान, ख़याल, खयाल	Incorrect
आधार	आधार	Correct
वर्ग	श्रेणी, दर्जा, वर्ग, कोटि	Correct

Table 5.1: Results obtained from the test document

This way we tested the system on documents from various domains. Table 5.2 summarises the results.

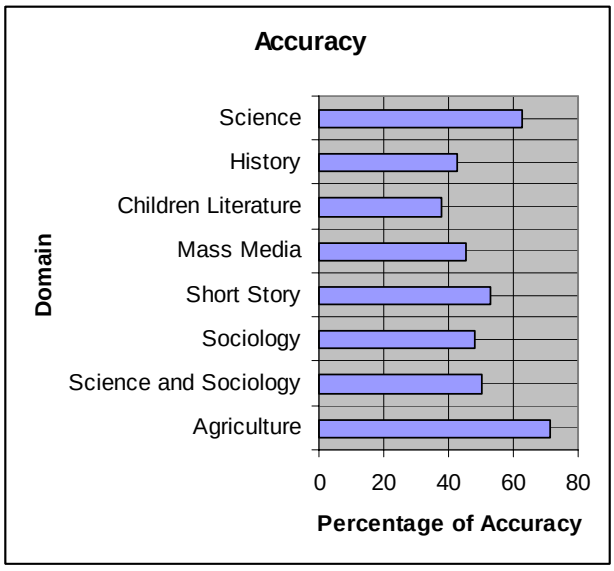

Figure 5.1: Histogram showing the WSD accuracy across domains for Hindi Words

Domain	Percentage of Accuracy
Agriculture	71.28
Science and Sociology	50
Sociology	48.34
Short Story	52.97
Mass-Media	45.45
Children Literature	37.78
History	42.85
Science	62.5

Table 5.2: WSD accuracy across domains for Hindi words

6. Conclusion and Future Work

In this paper we have used the Hindi Wordnet for a fundamental NLP task, *viz.*, disambiguation of Hindi words. To our knowledge, this is the first attempt at automatic WSD for an Indian language and is a significant step towards Indian language processing.

As can be seen, our accuracy values range from about 40% to about 70%. The performance can surely be improved if morphology is handled exhaustively. The system currently does not detect the underlying similarity in presence of morphological variations. Since Indian languages are rich in morphology, exhaustive pre-processing for morphology is crucial in the whole WSD process.

Our system currently deals with only nouns. Work is on to include words of other parts of speech. The obstacle there is the shallowness of the lexical network for non-noun words. With the enrichment of- for example, the verb hierarchy [18] - the system performance is expected to be very impressive.

7. References

[1] Fellbaum Christiane, editor the "WordNet: An electronic Lexical database". *MIT Press*, Map 1998

[2] Nancy Ide. "Parallel Translation as Sense Discriminators." *In Proceedings of SIGLEX99, Washington D.C. USA* 1999

[3] David Yarowsky "Word Sense Disambiguation using statistical model of Roget's categories trained on large corpora." *In Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454-460, Nantes, France, 1992

[4] Aggire E. and Rigau G." Word Sense Disambiguation using Conceptual density*" In Proceeding of COLING'96*.

[5] Adam Kilgrriff. "Gold standard Data-sets for Evaluating Word Sense Disambiguation Programs." In Computer Speech ansd Languages 12 (4), Special Issue on Evaluation, 1998.

[6] J. Pearl. "In Probablistic Reasoning in Intelligent Systems: Networks of Plausible Inference" Morgan K. Publishers, Inc.

[7] G. Ramakrishnan Deepa Prithviraj B. P. Bhattachryya S. Chakrabarti "Soft Word Sense Disambiguation,"*GWC-2003*.

[8] D. Chakrabarti D. Narayan P. Pandey P. Bhattacharyya "An Experience in Building the Indo-WordNet-A WordNet for Hindi." *GWC- 2002*.

[9] S. Jha D. Narayan P. Pande P. Bhattacharyya "A Wordnet for Hindi" Workshop on Lexical Resources in Natural Language Processing, India 2001

[10] D. Narayan and P. Bhattacharyya "Using Verb Noun Association for Word Sense Disambiguation" International Conference on Natural Language Processing (ICON 2002), Mumbai, India, December, 2002.

[11] C. Manning, H. Schutza – Foundations of Statistical Natural Language Processing *Word Sense Disambiguation Chapter*, The MIT press, Cambridge, Massachusetts, London, England.

.

[12] Dan Klein, Kristina Toutanova, H. Tolga Ilhan, Sepandar D. Kamvar, and Christopher D. Manning. "Combining Heterogeneous Classifiers for Word-Sense Disambiguation." In *Workshop on Word Sense Disambiguation: Recent Successes and Future Directions* at ACL 40, pages 74-80, 2002.

[13] H. Tolga Ilhan, Sepandar D. Kamvar, Dan Klein, Christopher D. Manning, Kristina Toutanova. "Combining Heterogeneous Classifiers for Word-Sense Disambiguation." *Proceedings of SENSEVAL-2, the Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pp. 87-90, 2001.

[14] Yarowsky, D. "Hierarchical Decision Lists for Word Sense Disambiguation" *Computers and the Humanities*, 34(2):179-186, 2000.

[15] Resnik P. and D. Yarowsky "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation" *Natural Language Engineering*, 5(2), pp. 113-133, 2000.

[16] D. Lin "An Information-Theoretic Definition of Similarity." *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin, July, 1998.

[17] M. E. Lesk "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone "Proc. *1986 SIGDOC Conference*, Toronto, Ontario, June, 1986

[18] D. Chakrabarti and P. Bhattacharyya, "Creation of English and Hindi Verb Hierarchies and their Application to Hindi WordNet Building and English-Hindi MT*"*, (GWC-2004), Czech Republic.

[19] Hindi Wordnet from Center for Indian Language Technology Solutions, IIT Bombay, Mumbai, India http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[20] Hindi Corpora from Central Institute of Indian Languages, Mysore India. http://www.ciil.org

Acknowledgement

The research is supported by a grant from "Ministry of Information Technology and Communications, Government of India, New Delhi" and "Develop Gateway Foundation".

Appendix - I

Hindi Wordnet API used for WSD.

- **char * morphstr (char *origstr, int pos):** Finds the base form of the word *origstr* (original string) in the specified *pos*. The first call (with *origstr* specified) returns a pointer to the first base form found. Subsequent calls requesting base forms of the same string must be made with the first argument of NULL. When no more base forms for *origstr* can be found, *origstr* itself is returned.
- **unsigned int in_wn (char *searchstr):** Finds the part-of-speech. Returns an unsigned integer with a bit set corresponding to each syntactic category containing *searchstr*. 0 is returned if searchstr is not present in *pos*.
- **IndexPtr index_lookup (char *searchstr, int pos)**: Finds *searchstr* in the index file for *pos.* Returns a pointer to the parsed entry in an Index data structure. Returns NULL if a match is not found.
- **char * findtheinfo (char *searchstr, int pos, char *ptr_type, int sense_num):** Searches the database for relational information of a word. Returns a pointer to the text buffer. *ptr_type* gives the pointer to the relations in Wordnet and *sense_num* is the particular sense number.