Word Sense Disambiguation Using Hindi WordNet and Lesk Approach

Manisha Gupta¹, Seema Yadav², Shraddha Sharma³ and Dr. Surendra Yadav⁴

¹, ², ³ Department of Computer Science, Banasthali University, Jaipur (Rajasthan)
⁴Professor and Head of CS Department, Maharishi Arvind College of Engineering & Research Centre, Jaipur

ABSTRACT

Word Sense Disambiguation (WSD) is a vital area which is very useful in today’s world. WSD is the process of automatically clarifying the meaning of a word in its context. There are many words which has multiple meanings. So there is a need for disambiguate these words with the correct sense. In this paper we use Hindi Language which is the first language of the Indian Populations. We will disambiguate the Hindi words. In this paper we use Hindi WordNet for multiple senses of a word and Lesk Approach to disambiguate the correct sense of the word. So in this paper we take the so-called Lesk Approach. In our case, definitions of the senses of the words to be disambiguated, as well as of the ten surrounding nouns, verbs, adjectives, are derived and enriched using the Hindi WordNet Lexical Database. In this we use modifies Lesk Approach.

Keywords: Word Sense Disambiguation, WSD Approaches, Hindi WordNet, Lesk Algorithm.

1. INTRODUCTION

Language is a communication media for the creatures among the races. The language can able to exchange the information among the races. Language is the evolution criteria for the technology, the development of the Languages can able to exchange the thoughts, views, suggestions in an understandable way. The development in the human races has been observed when the communication among the people started to increase from the rock age [1]. In this paper we use Hindi Language. Hindi Language is the First Language for the Indian People. Most of the Indian People use Hindi Language in their communication. So there are also ambiguities in the Hindi Words.

1.1. Word Sense Disambiguation:-

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. In Natural Language Processing (NLP), Word Sense Disambiguation (WSD) is a common problem which identifies the sense of the word used in the sentence or the query when it has multiple meanings [2]. WSD is used to find the correct meaning of the sense or the word. A word can have number of senses, which is termed as ambiguity. This word sense disambiguation is an intermediate task, but rather is necessary at one level to accomplish most natural language processing tasks.

A rich variety of techniques have been researched from dictionary-based methods that use knowledge encoded in lexical resources, supervised machine learning works on classifiers and unsupervised learning method supports clusters and many more as such [3]. In this paper we have some words with the more than one meaning. In this we use our Corpus. In this corpus we have 10 to 15 sentences. There are many approaches which use Corpus based methods. Some approaches uses training data and some employ external data. Different approaches to word sense disambiguation have been taken. Many are based on different statistical techniques. Some require corpora that are tagged for senses and others employ unsupervised learning. In this paper we take the so-called Lesk approach (Lesk 1986), which involves looking for overlap between the words in given definitions with words from the text surrounding the word to be disambiguated.

In real life there are many words which are Polysemous in nature for Humans and Computers. So this is a rarely problem not only for humans but also for computers.

Ambiguity for Humans: - Ambiguity for humans is the problem in day to day communication.
Ex:- lhrk us dgk fd jkqgq eka ds lkFk ctkkj x;kA
In this sentence there are two senses for the Humans.
(\k rks iSny x;k ;k xkMh ij x;k)  
(\k rks jkqgq viuh eka ds lkFk x;k ;k lhkr dh eka ds lkFk x;k)
Ambiguity for Computer: - Ambiguity for computers is also rarely problem in day to day communication. 

Ex: -lksuk lksuk pkgrh gSA
In this sentence the word “lksuk” has three meanings.
1. “Name of a Girl”.
2. ” Gold”.
3. “Sleep”.

As the human use its sense for finding the correct meaning of a word in the sentence he is able to disambiguate the meaning of a word manually but problem occurs for a machine because a machine could not have that common sense by which it can conclude the appropriate meaning of an ambiguous word [4].

In this paper we discuss ambiguity for computers.

2. HINDI WORDNET

Word net is a network of words linked by lexical and semantic relation. Word net is a large lexical database of English noun, verbs, adjective, and adverb. They are grouped into sets of synonym or synsets. The Hindi WordNet is developed by Prof. Pushpak Bhattacharyya, Centre for Indian Language Technology (CFILT), Computer Science and Engineering Department, IIT Bombay [5]. The Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. In the Hindi WordNet the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Hindi WordNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNet. Thus, the Hindi WordNet contains the following category of words- Noun, Verb, Adjective and Adverb [6].

Each entry in the Hindi WordNet consists of following elements: -

2.1. Synsets: - It is a set of synonymous words. For example, “फल, परिणाम, अंजाम”

2.2. Gloss: - It describes the concept. For example, “किसी कार्य के अंत में उसके फलस्वरूप होनेवाला कार्य या कोई बात”

For Example: - “उसके कार्य का परिणाम बहुत तही बुरा लिखला”

In the Hindi WordNet there are:
Total number of synset: 28,687
Total number of unique words: 63,800

There are many semantic relations in Hindi WordNet Which are: -

<table>
<thead>
<tr>
<th>Relation</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypernymy/Hyponymy</td>
<td>Is-A (Kind-Of)</td>
<td>बेलपग Is A Kind-OF पत्ता</td>
</tr>
<tr>
<td>Entailment/Troponymy</td>
<td>Manner-Of (for verbs)</td>
<td>खरा[टा लेना, नाक बजाना==&gt; सोना</td>
</tr>
<tr>
<td>Meronymy/Holonymy</td>
<td>Has-A (Part-Whole)</td>
<td>गड ं Is The Part-OF पेड</td>
</tr>
<tr>
<td>Antonymy</td>
<td>holds between two words that express opposite meanings</td>
<td>मोटा =&gt; पतला</td>
</tr>
<tr>
<td>Causative</td>
<td>convention of forming causation</td>
<td>खाना =&gt; खिलाना</td>
</tr>
</tbody>
</table>

3. LESK APPROACH

In this paper we use Lesk Algorithm (Which is implemented by Michael Lesk 1986: Identify senses of words in context using definition overlap) for disambiguation the words. Lesk Algorithm is an Overlap Based Approach. Overlap Based Approaches generally require a Machine Readable Dictionary (MRD). The determination between the features of different senses of an ambiguous word (sense bag) and the features of the words in its context (context bag) [3].
Lesk algorithm is explained as follows:-
Consider an ambiguous word W and let C is the collection of a set of context words in its surrounding window. Search the word W in the Hindi WordNet and find the senses of that word W. So, there will be lot of senses S for W of words. Now we use Lesk algorithm to find the best sense of the word W. So, at last the output will be the sense as the most probable sense which has the maximum overlap [1]. Here, we represent Lesk Algorithm which is implemented by Michael Lesk in 1986. As if there is more than one word in the sentence then we can remove ambiguity with simulated annealing.

function LESK (word, sentence) returns best sense of word

1. consider the function LESK (word, sentence) returns best sense of word.
2. best-sense = 0
3. max-overlap = 0
4. context = set of words in sentence
5. for each sense in senses of word do
6. content = set of words in the gloss and examples of senses
7. overlap = COMPUTEOVERLAP (content, context)
8. if overlap > max-overlap then
9. max-overlap = overlap
10. best-sense = sense

end return (best-sense)

4. GLOSSES FROM HINDI WORDNET

In our experiment if we find senses for the word “lksuk” then we use Hindi WordNet. There are many senses for the word “lksuk”.

For an example if we search the correct meaning for this sentence: - “jke ds ikl lksus dh ekyk gSA”
So, senses for the word “lksuk” in Hindi WordNet are:-
Query: “lksus dh ekyk”

1. lksuk : As Noun

    Sense 1 : "k'u dh fdzzz;k ("k'u ds fy, jkr cukbZ xbZ gSA)
    Sense 2 : "d cqgqeq'; ihy /凶eqlftlds xgus cers gSA (вktdy lksus dk Hkko vkleku Nq jgs gSA)
    Sense 3 : "d izdkj dk isM+ (lksukikBk ds cht Nky vksj Qy nok ds :i es dke vkr$s gSA )

2. ekyk : As Noun

    Sense 1 : ysvDj "kjij vksj efLr"d dks foJke nsus dh voLFkk ("k'u ds fy, gh jkr cukbZ xbZ gSA)
    Sense 2 : jDr lapkj ;ds;j iGkFk ;k iSj ds fdih Hkkx dk IqUu gksuk (jkr ds le; esjk gkFk Ihus ds uhps ncs ds dkj.k lks x;kA)

So, the word “lksuk” has 5 senses in the Hindi WordNet. If we connect that word “lksuk” with the word “ekyk” then there will be different sense for the sentence “lksus dh ekyk”.

Senses for the word “ekyk” in Hindi WordNet are:-

2. ekyk : As Noun

    Sense 1 : lwr es fijkbZ xksykdkj vkdfr dh oLrq tks ,d xgus dh rjg xys es iguh tkrh gS (melds xys es ekfr;ksa dh eky lq'kksfHkr gks jgh gSA)
    Sense 2 : cgqeeqyhykjkse es AjL uhus ds fopkj ls cus edku dk Lrj (esjk ?kjIk;roh eafty ij gS )
    Sense 3 : sih iEijk ftilsa ,d gh izdkj dh oLrq,a, O;fDr ;k tho ,d nwls ds ckn ,d lh/k esa gksa (jk'ku dh nqdku ij yksxksa dh iafr yxh gq$bZ gSA)

So, the word “ekyk” has 3 senses in the Hindi WordNet.
At last by The Lesk Algorithm the best sense of the word “ekyk” is “xguk”.

In this sentence there is an overlap word which is “xguk” so the result we get is “xguk”. In this we avoid stop words which are (dk, ds, dh, gS, gks, ls, esa, us etc.)

If we want to disambiguate the word “Qy”.

Let, the next example is:-

“Lhrk dh ijh[kk dk Qy vk x;kA”

So, senses for the word “ijh[kk” in Hindi WordNet are:-

Query: “ijh[kk dk Qy”

1. ijh[kk : As Noun
   Sense 1 : ;ksX;rk lkeF;Z] xq.k vkfn tkuus ds fy, vPNh rjg ls ns[kus ;k ijjkus ds fdz;k ;k Hkko (rzkS.kkp]Z us vtqZu dh ijh[kk ds fy, eNyh dh vkj;k es rjh ekjus dks dgk )
   Sense 2 : fdlh dh ;ksX;rk ;k Kku dks ijjkus ds fy, mls iz”u iwNus dh fdz;k ftlds vk/kkj ij mldk ifj.kke fudkyk tkrk gS (lhrk nloha dh ijh[kk es vPNs vad ykus ds fy, th rhsM+ esgur dj jgh gS )

So, there are 2 senses for the word “ijh[kk”.

If we connect the word “ijh[kk” with “Qy”. Then there will be different meanings.

So, the senses for the word “Qy” are:

2. Qy
   Sense 1 : ouLifr esa gksus okyk xwns ;k cht ls Hkjiwj chtdks’k tks fdlh fof”k”V _rq esa Qwy vkus ds ckn mRiUu gksrk gSA (vke lcls ehBk Qy gS)
   Sense 2 : fdlh dk;Z ds var esa mlds QyLo:; k gksus okyk dk;Z (cqjs dk;Z dk ifj.kke Hkh InSo cqjg gh gksrk gS)
   Sense 3 : ;d lqxaf’kr Qy tks vksk”k/kh vkSj elyks d dhe esa vkrk gS (tk;Qy dks f?ldj cPjkSa dks [kkjlh vkus ij f;y;k tkrk gS)
   Sense 4 : rjg ;k cjNh vkfn ds rst /kkjokyk ;k vksxs dk /kkjnkj Hkkx (bl rjg dk Qy cgqr uqfdyk gS)
   Sense 5 : ryojk vkfn dk okj jksdus dk ,d midj.k (<ky ,ks/nkvksa dks lqj[kk iznku djrh gS)
   Sense 6 :fdlh dks m/kkj fn;s gq ,;k cSad vkfn esa tek fd, ;i;ksa ds cnyr esa ml le; rd feuys okyk og fuuf”pr /ku ftp le; rd ewy /ku
      okil fey u tk, (” k us C;kt ij ;i; yxk j[jks gS ftldk mls ;d vPNk Qy izkIr gksxk )
   Sense 7 : Qyfr T;ksf;r” esa xzgksa dh FLFkfr vkJ] ;ksx dk ifj.kke tks fdl nj[k ;k lqj[k ds ;i esa gksrk gS (cgqr yksx jkf”kQy es fo”okl jk;[ks gA) 
   Sense 8 : Qy ;k vadks vkfn ds ;i esa izklr og ifj.kke ftldk fy, xf.kr dh dksbZ fdz;k dh tkrh gS ((ks=Qy xq.kQy ;ksxQy HkxQy tkf ifj.kke) ds ;i esa ifj.kke)

In this there are 8 senses of the word “Qy”.

So, the correct sense of the sentence “ijh[kk dk Qy” is “ifj.kke”. Here also we can avoid stop words.

So, by the Lesk Algorithm we can search the correct sense for the ambiguous words. We have many examples for the disambiguation like:-

1. vad ,ksx
2. x;k eas tkuk
3. yksxksa dk ny
4. lksuk pkguk
5. n.M CSBd
6. xqykc dh dye

5. RESULT AND ANALYSIS

The Lesk Algorithm is crucially dependent on the length of glosses. This algorithm is based on the perception that the words that co-occur in a sentence are being related to a same topic so that their senses can provide a gloss which can give an overlapped word which is able to define the best sense of ambiguous word or the target word. A word can be disambiguated by finding that sense whose definition shares the most number of words with the definition of the neighbouring words of same sentence. As in query the glosses of “lksuk” and “ekyk” gives the sense of “lksus dh
The algorithm is using a supervised approach the system is trained before applying the algorithm. This approach typically makes use of other sources of information for example Lesk Algorithm uses the info contained in a dictionary to perform Word Sense Disambiguation. This algorithm is based on intuition that sentences being used are somehow related so they can provide related senses of words. For example if the words “jkh[kk]” and “Qy” occur together in sentences one may presume that their intended senses both refer to same topic. These two senses should be defined in a dictionary using some of the same words. So that the actual sense from the overlapped words can be obtain. This word can be disambiguated by finding that sense whose definition shares the most number of words in same sentence. Since the algorithm is dependent on finding common words between definition it suffers from the fact that lexicographers generally aims at creating concise definitions with as few words as possible. Lesk algorithm restricts its comparison to just the dictionary meanings of words. In future the Lesk algorithm can be successfully adapted to Hindi WordNet to produce an unsupervised word sense disambiguation system. This system will be reasonably accurate compared to other systems. A local disambiguation algorithm can be made using a heterogeneous scheme. Instead of finding overlapped words among only the glosses of words a richer source of information can be used. This heterogeneous scheme include hierarchy built on the hypernym, holonymy, metonyms, Troponymy, Is a part of, and Has a part of links etc. Not only the glosses of words but also the glosses of their synsets can also be utilized. Although the data will become redundant but this approach will definitely improve disambiguation accuracy of Hindi words.

REFERENCES


ABOUT THE AUTHORS

Manisha Gupta is pursuing her M.Tech degree in Computer Science and Engineering from Banasthali University, Jaipur, Rajasthan and received B.Tech degree in Computer Science and Engineering from Rajasthan Technical University, Kota in 2011. Now, she is doing research in Natural Language Processing from Banasthali University, Jaipur.

Seema Yadav is pursuing her M.Tech degree in Information Technology from Banasthali University, Jaipur, Rajasthan and received B.Tech degree in Computer Science and Engineering from Rajasthan Technical University, Kota in 2011. Now, she is doing research in Natural Language Processing from Banasthali University, Jaipur.

Dr. Surendra Yadav, Professor and Head of Computer Science Department in Maharishi Arvind College of Engineering & Research Centre, Jaipur. He has done his Ph.D 2012 form Rajasthan University. He has done his M.Tech in 1998 from BIT Ranchi.