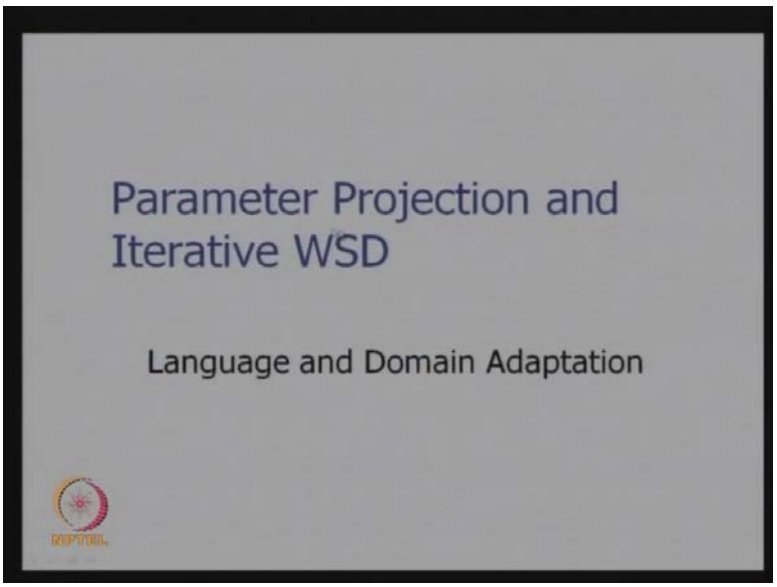
Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science & Engineering
Indian Institute of Technology, Bombay

Lecture No - 36
Resource Constrained WSD; Parsing

What we have done in the last lecture is that we have complete a discussion of supervised, semi supervised and unsupervised words and disambiguation. We also mention that we will briefly touch upon resource constraint to words disambiguation. Unsupervised words and disambiguation was a method to do words and disambiguation, when one does not have training corpus, training corpuses is a costly resource which takes time money and manpower to create. Now, there are many languages which were resource constrained, they do not have training corpus; they a do not have n l p tools, techniques and so on. So, we have a investigated the problems of resource constrained words and disambiguation in IIT Bombay, and I would like to briefly discuss this.

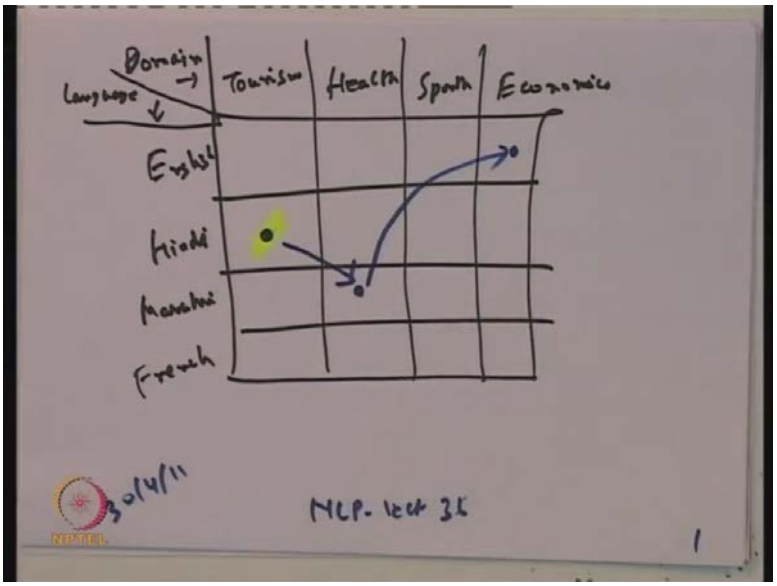
(Refer Slide Time: 01:17)



So, this kind of a work is called a parameter projection, and iterative words and disambiguation, parameter projection is based on making use of multilingual coordinate, where since x are aligned and these ideas are useful for language and domain adaption, the meaning of language and domain adaption is that, if we have words and disambiguation in 1 language, when the work can be reused for another language.

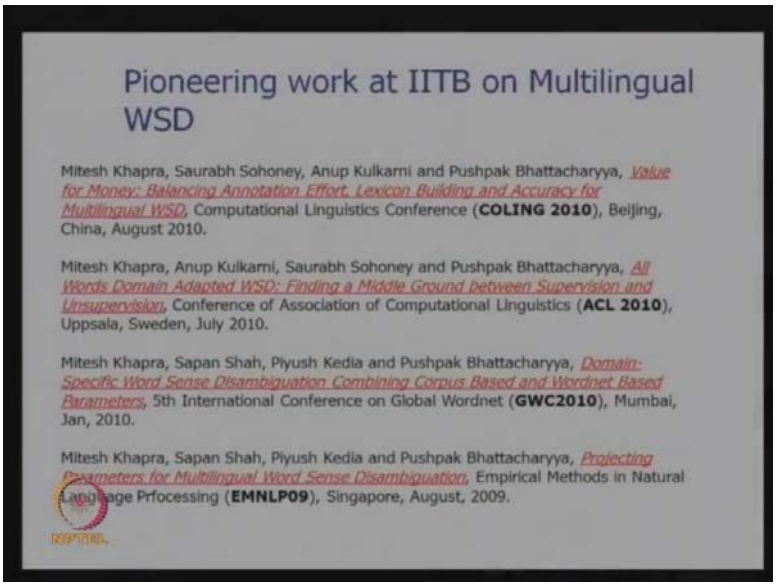
Similarly, if we have words and disambiguation for one domain said tourism, we can re use the work for another domain, say hint. So, language at domain adaption is a very important idea, and I would like to present a matrix, which is drawn on the paper, to say the immense potential of the work.

(Refer Slide Time: 02:25)



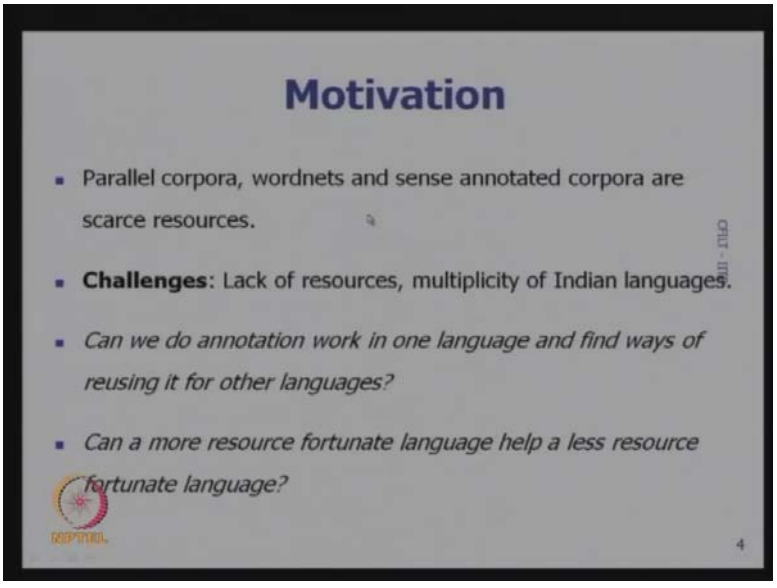
So here is a matrix where we have domain and Language so we can have things like, tourism, health, sports, economics, and so on. And here, we have languages, let us say English, Hindi, Marathi, French and so on. So, if we have domain and Language adaption, we suppose, we have created a words and disambiguation system here, this is the words and disambiguation system, built for Hindi and Tourism. Now, is possible to make use of this work, and create words and disambiguation for let us say, Marathi and health, here can we do this. So, we adapt for the system built for Hindi, and create a system for Marathi and Health, from this we can say, we can go to a system for English and Economics. So, this is an very interesting possibility, just like a virus or a team of ants produces a colony, spreading from a center, and creating the colony around it. Similarly, our vision is that, we have, a words and disambiguation before 1 Language and 1 Domain. And its spreads around for different languages domains, this is the idea. So, can we do this now looking at the slides?

(Refer Slide Time: 04:47)



We have done a number of interesting, work on a multilingual words and disambiguation. So, they were presented into top forum like COLING ACL, empirical methods in natural language processing, global learning conference and so on. So, we proceed to discuss the techniques.

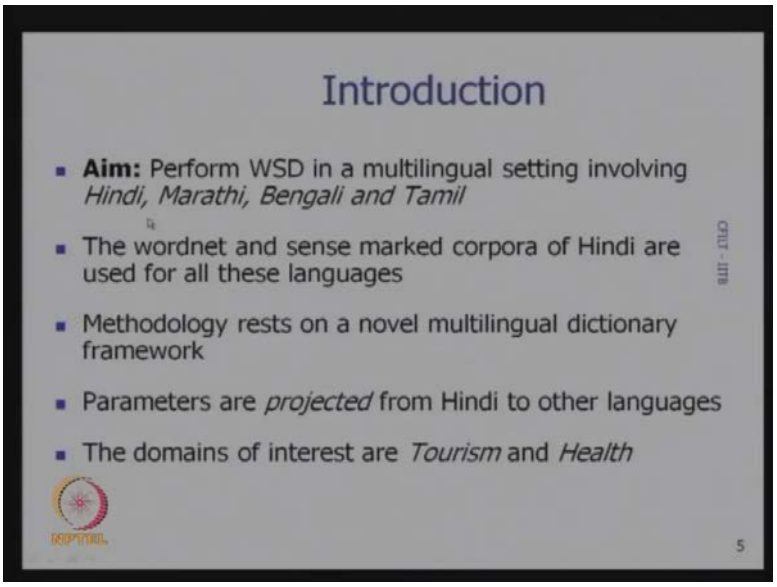
(Refer Slide Time: 05:14)



Before that the motivation, as has been said parallel corporate of wordnets and sense annotated corpora, are scarce resources, the challenges are lack of resources multiplicity of Indian languages. So, can we do annotation work in language, and find ways of

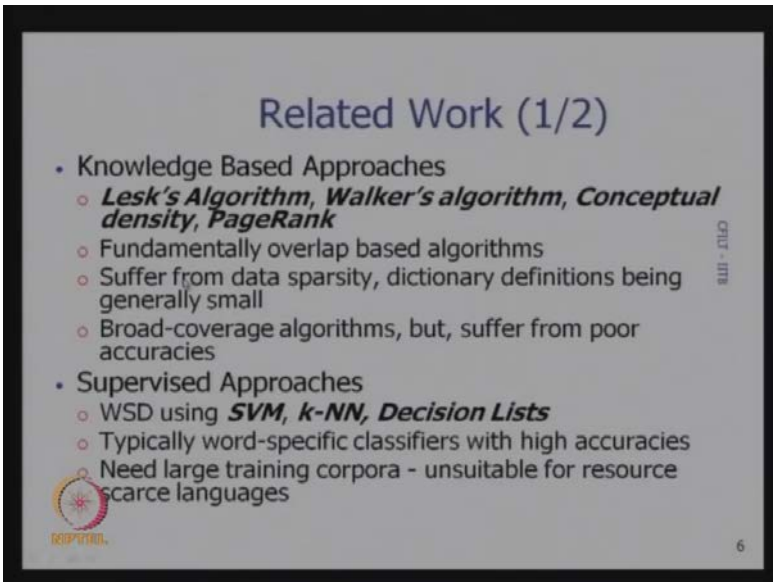
reusing if for other language, can a more resources fortunate language help a less resource fortunate languages.

(Refer Slide Time: 05:40)



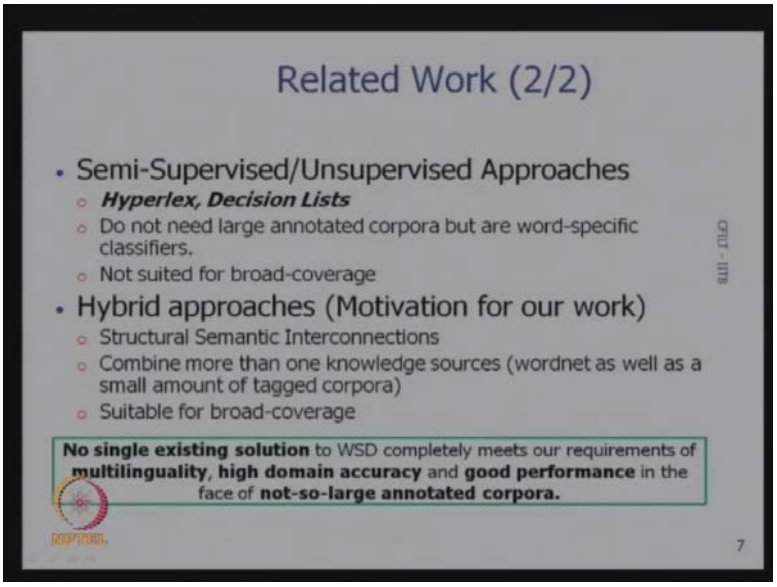
So, the introduction is that, we would like to perform WSD in a multilingual setting involving Hindi Marathi Bengali and Tamil, the wordnet and sense marked corpora of Hindi are used for all these languages, the methodology rested on a novel multilingual dictionary framework parameters are projected from Hindi to other languages, the domains of interest are Tourism and Health.

(Refer Slide Time: 06:03)



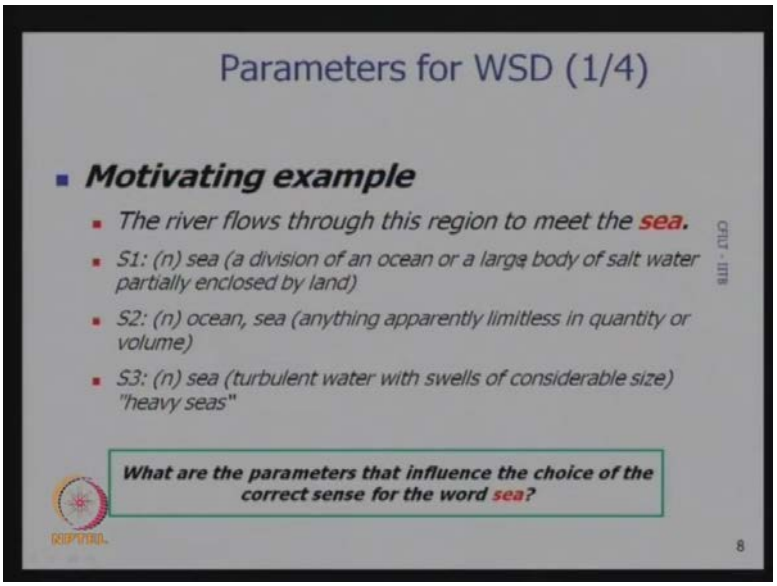
So, this line work derives a its base ideas and inspiration from well known algorithms, Lesk's algorithm, Walker's algorithm, Conceptual density, Page rank based algorithm which are knowledge based approaches them SVM, k-NN neighborhood decision lists which are based on supervised approaches. It also draws from semi supervised to un supervised approaches.

(Refer Slide Time: 06:23)



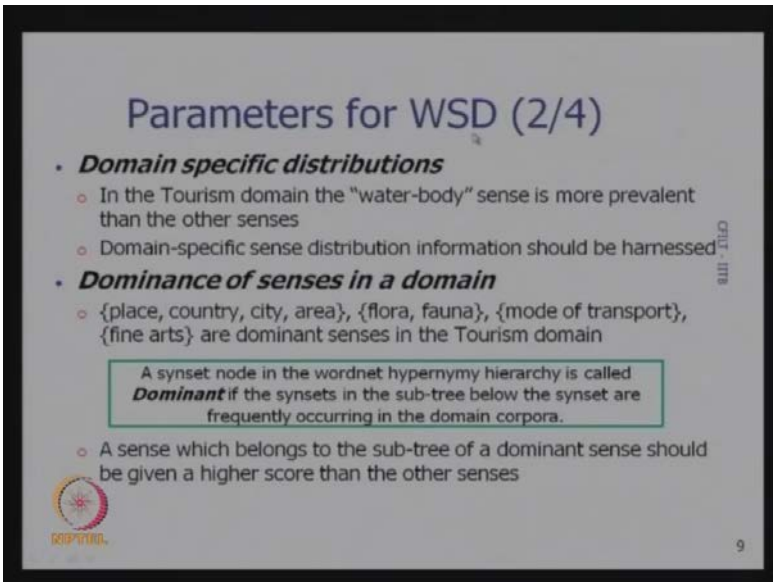
Like hyperlex, decision lists, and hybrid approaches are the motivation for our work, structures semantic interconnections, combine more than one knowledge source, wordnet as well as small amount, for broad coverage so this is suitable for broad coverage. Now, here is the punch line of the work. No single existing solution to words and disambiguation, completely meets our requirements of multilinguality, high domain accuracy, and good performance in the phrase of not so large generative corpora. So, before we begin our discussion, we first identify the crucial parameters for words and disambiguation as has been said many times, when a word is disambiguated from the neighboring words. So, that naturally creates a setting or parameters of words and disambiguation clearly, if you are using training corpus, then the sense of words, which occurs most frequently, will be a very important parameter for disambiguation. The relationship of this word, with the words in the neighboring context and their senses is also very important. So, these considerations give rise to a number of parameters, which are shown on the slide.

(Refer Slide Time: 08:05)



Here is a motivating example, the river flows through this region to meet the sea, so the target word for disambiguation is sea. The meanings of sea are given here as obtained from the English wordnet, sea as a noun, a division of an ocean or a large body of salt water, partially enclosed by land, sea can mean anything apparently limitless in quantity or volume. This is metaphoric uses other meaning of sea is turbulent water with swells of considerable size, heavy seas is their example. So, what we see is that, there are 2, so called physical senses of sea, in the sense of water body, and a metaphorical, senses of anything limitless the quantity or volume.

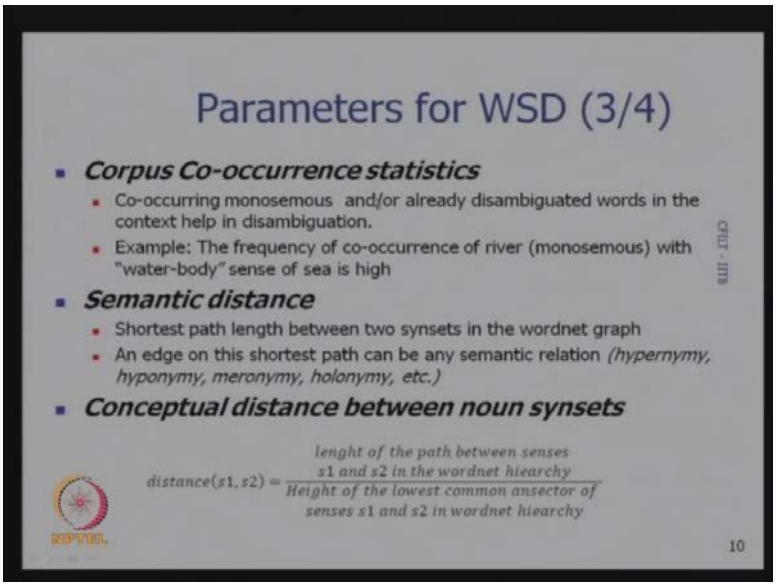
(Refer Slide Time: 09:01)



Now, we will see that domain specific distributions, for an important component of the parameters for WSD domain industry. In the Tourism Domain, the water body sense is more important prevalent than the other senses, domain specific sense distribution information should be harnessed. So, this is the our first parameters, coming from the corpus sense of corpus, dominance of senses in domain that is a very important parameter place, country, city, area, flora fauna mode of transport, fine arts are dominant senses, in the tourism domain. So, what we see is that, a synset node in the wordnet hypernymy hierarchy is called dominant.

If the synsets, in the sub tress below the synset, are frequently, occurring in the domain corpus. So, dominance is determine not only by the frequency of a particular sense but also, its children in the hypernymy hierarchy is descendants hierarchy. So, the sense being dominant is conditioned on a, on pretty stringent criteria its own frequency, and also its descendants frequency, a sense which belongs to the sub tree of a dominance sense, should be given a higher score than the other senses.

(Refer Slide Time: 10:25)



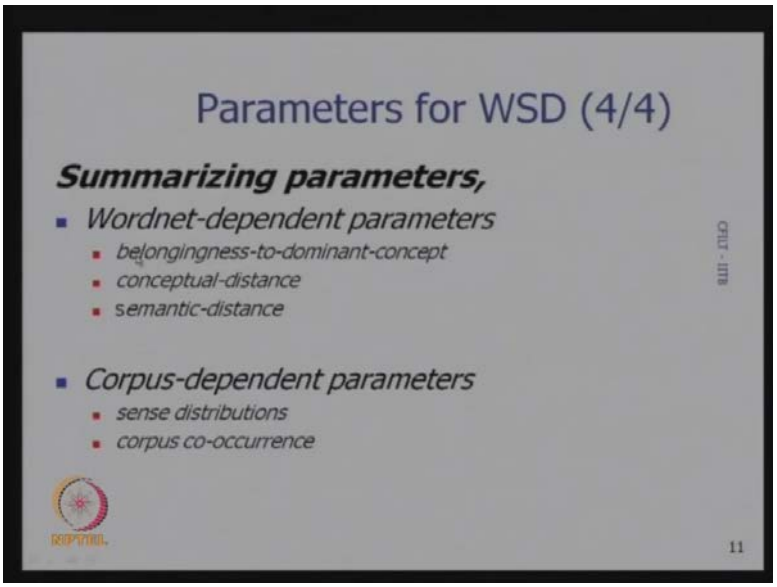
Corpus co-occurrence statistics is a very important, co-occurring monosemous and or already disambiguated words in the context help disambiguation, for example, the frequency of co-occurrence of river, which is monosemous with water body, sense of sea is high. So, the sea also has its metaphorical sense of something limitless, but that will not have high co-occurrence, that sense will not have high co-occurrence, with river.

River is a monosemous word by the way, semantic distance is the shortest path length between 2 synsets in the wordnet graph, an age on this shortest path can be any semantic relation.

Then there is these notion of conceptual distance between, noun senses the distance between 2 senses s 1 and s 2 is defined as the length of the path between the senses s 1 and s 2 in the word net hierarchy, divided by the height of the lowest common ancestor of senses s1 and s2 in word net hierarchy. So, this particular point is when we mention the conceptual distance, words and disambiguation algorithm. So, the distance between 2 senses is important, but distance has to be seen, in the context of how general distances are, we had mention before, that he distance between cat and dog, which are children of the animal node is 2, the cat and dog are siblings of each other.

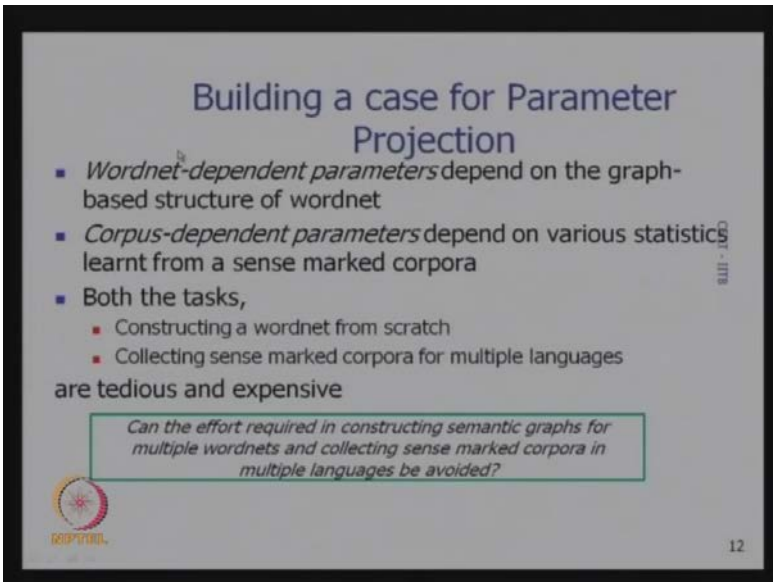
Similarly, the concepts abstract, and concrete are also siblings of each other. But they come very high in the conceptual hierarchy, clearly the distance, the code on code distance between abstract and concrete which is very large concepts even though the length is 2 between them that is not representative of actual conceptual distance between abstract and concrete. Cat and dog are very similar, but abstract and concrete are 2 large conceptual words. They are very distant from each other, in terms of semantics and concept conceptual categorization. So, the height of those conception in conceptual has an hierarchy is important.

(Refer Slide Time: 13:09)



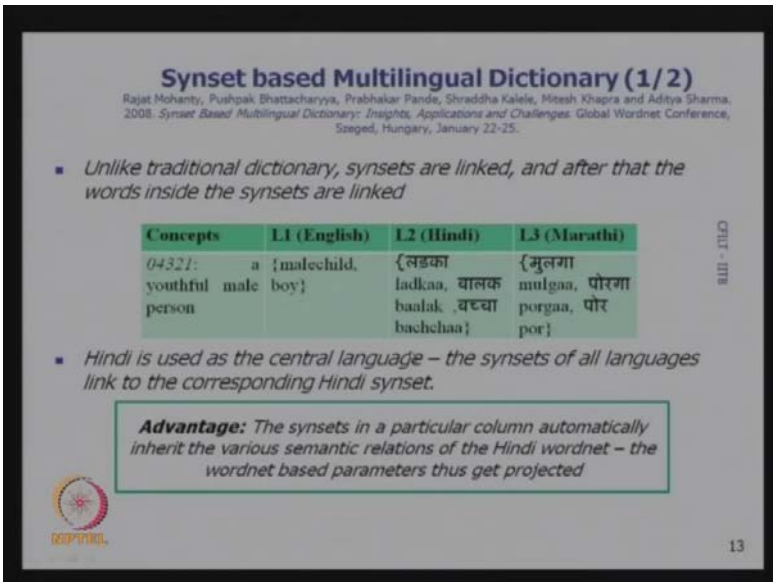
There are other parameters which are there but the mean parameters for wordnet dependent parameters, belongingness to dominant concept, conceptual distance and semantic distance, they come from the wordnet. And corpus dependent parameters are sense distribution corpus co occurrence distribution, they come from the corpus.

(Refer Slide Time: 13:32)



Now, we would like to build a case for parameter projection, wordnet dependent parameters, depend on the graph based structure of wordnet, corpus dependent parameters, depend on various statistics, learnt from a sense marked corpus. Both this tasks requires a constructing a wordnet from scratch, collecting a sense marked corpora from multiple languages. And these tasks are tedious and expensive, the question we are asking is can the effort required in construction semantic graphs, for multiple wordnets and collecting sense marked corpora, in multiple languages be avoided.

(Refer Slide Time: 14:10)

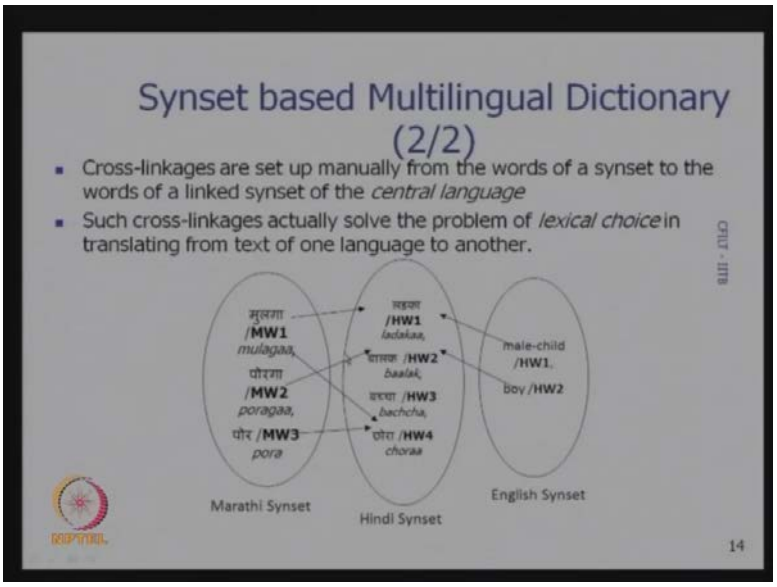


So, these ideas are based on a novel, synset based multilingual dictionary, this was published in global wordnet conference 2008. The authors are mentioned here Rajat Mahanty, me, myself Pravaker pandey, Shraddha and so unlike tradional dictionary synsets, are linked and after that the words, inside the synsets are linked. If we look at this 1 row from the dictionary, the multilingual dictionary, this row is the concept of youthful male person; this is expressed by the word boy and male child in English. This concept is expressed by the word [fl] in Hindi; these are a also expressed by the words [fl] in Marathi. Now, Hindi is also used as a private language, the synsets of all languages link, to the corresponding Hindi synset. The advantage of this is this representation is that, the synsets in a particular column automatically inherit, the various semantic relations of Hindi wordnet.

The wordnet based parameters thus get projected, so if we look carefully, we see that each column in that multilingual dictionary is nothing but the word net of that Language. So, this whole column is a English word net, this is Hindi word net, this is Marathi word net, and all the synsets ids are placed in alignment, to each other. Then within a particular row, the words are linked to each other for example, boy and linked to [fl], and also it can be linked to [fl] but these things capture, what is called the register. Now, suppose the setting is informal, then the word boy linked with the word [fl], with less weightage, because maybe the discourse setting is that the informal sense of a boy, is not very

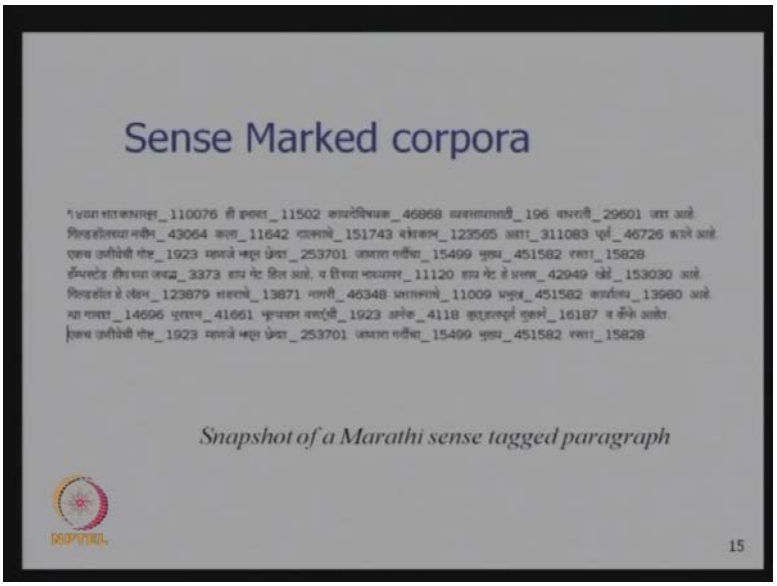
frequent. So, the main point being here is that the concepts are linked and, within the concepts the words are crossed linked internally.

(Refer Slide Time: 16:47)



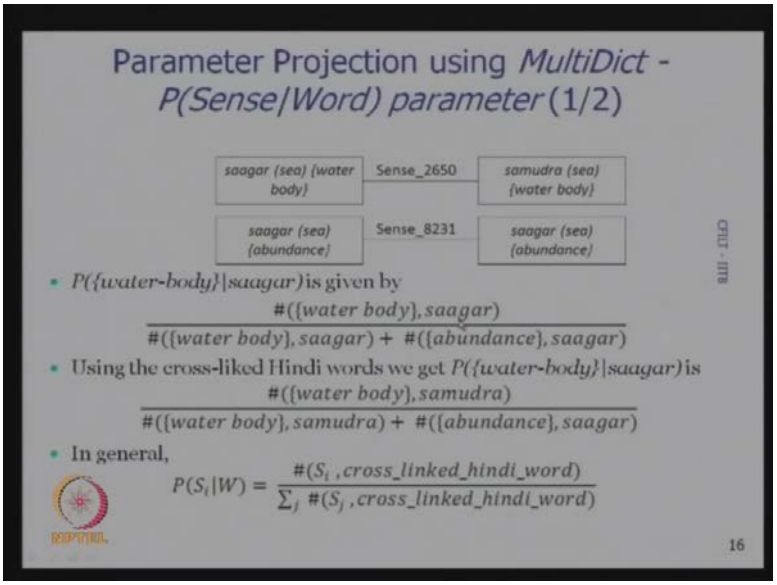
The same notions are expressed here, if we take the Marathi words [fl] them [fl] is a informal word in Marathi [fl] is an informal word in Hindi. So, a link has established from [fl] the word [fl] in [fl] are more formal and they are linked to the word [fl] and balak for English the word boy links to [fl] and [fl]. So, this is a novel concept based dictionary, where the concepts are linked and within them, the words are linked. This produces a very rich lexical network, the concepts are linked that within them or linked. So, this produce, this becomes a very rich structure of word link ages. And from it one can very easily create, programmatically a bilingual dictionary, once these, rich structure is placed, various mappings can be established between the words.

(Refer Slide Time: 17:59)



Proceeding further, these kinds of sense marked corpus, is used for words and disambiguation. The training of the system words and disambiguation is done by means of these kind of sense marking. And here, we see that a word net id's are placed alongside the words [fl] for example, has many senses but in the context that is shown the id's 11502. This is a particular sense of [fl] in the sense of art Kala, in the sense of art is highly polish us, in this particular context the id is 11642. So, this is the way, the sense for corpus is built, and this example comes from Marathi.

(Refer Slide Time: 18:58)



Now, we take an example to understand the concept parameter projection, using a multidict, the parameter probability of the sense given in the word is a highly important parameter, this is the first parameter under discussion. One can see that this particular has to come from, sense mart corpus, we have to count, the frequency of the sense, given in the word. Let us take the word sea or let us take the word sagar in Hindi, which has these 2 senses water body and abundance. The second sense is metaphorical [fl] sea of knowledge, this sense id is 2650 and it maps to the word [fl] water body sense in Marathi. The abundance sense is link to the word [fl] in Marathi again with the sense of abundance; the sense id is same in the both cases 2650 for the water body sense.

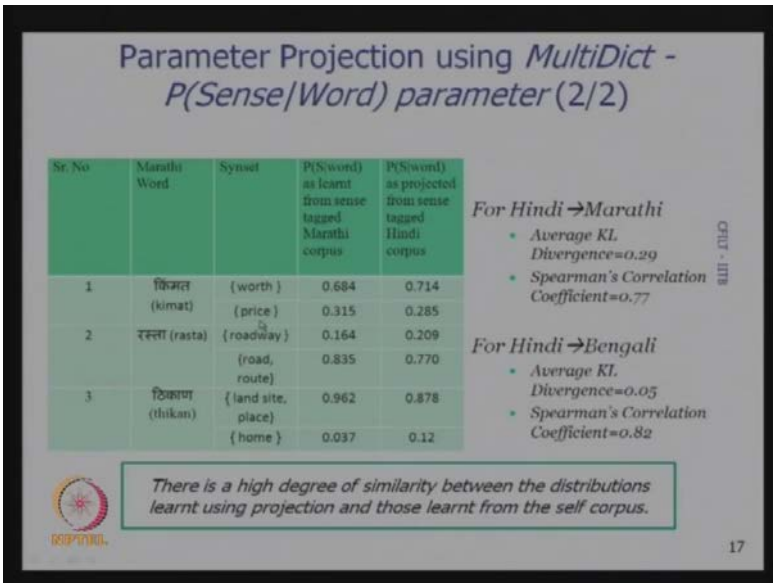
Both in Hindi and Marathi, the sense id is 82311 in Hindi and Marathi abundance sense id to the word [fl] in Marathi, and with the sense of abundance the sense id is same in both cases 2650 for the water body sense, both in Hindi and Marathi the sense id is 8231 in Hindi and Marathi for the metaphorical sense of [fl]. Now, how will the probability water body sense given [fl] will be calculated? They will be calculated by a frequency count, we will see how many times [fl] appeared with the sense of water body. These will be divided by the number of times the word [fl] appears assuming [fl] has only 2 senses water body in abundance. And if we take all possible senses of [fl] this will be exactly equal to the number of times word [fl] appears in a corporate. So, water body sense of a [fl] in the numerator, the number of times that happens divided by the number of time [fl] appears with water body, and the number of times [fl] appears with abundance.

This sum is equal to the number of times the word [fl] itself appears. Now, if we use the cross linked Hindi word, we get P water body given [fl] as water body and [fl] appear together, divided by water body and [fl] appearing, and [fl] and abundance appearing together. So, in general, the probability of a sense given in the word, of a different Language, different from Hindi, is equal to the sense, and the crossed linked Hindi word appearing together, divided by different senses with cross linked Hindi word, appearing with the corpus. So, I suppose it is clear what we are discussing, we want to calculate the water body sense of [fl] in Marathi but we do not have training corpus of Marathi, which is sense marked.

Suppose, now this will be equal, to the water body sense of [fl] divided by the number of times [fl] appears or divided by the number of times [fl] appears as water body and [fl] appears as water body and [fl] appears with the sense of abundance. So, we will obtain,

the count from the crossed linked Hindi word, we will see the cross linked Hindi word for [fl]. And we will see in how many times; we will see how many times if appears in the sense of water body, in the Hindi training corpus, how many times does it appears in the abundance sense, in the Hindi sense of corpus. And from this, we will be able to estimate, the probability of the Marathi word, in a particular sense, now does this procedure work.

(Refer Slide Time: 23:55)



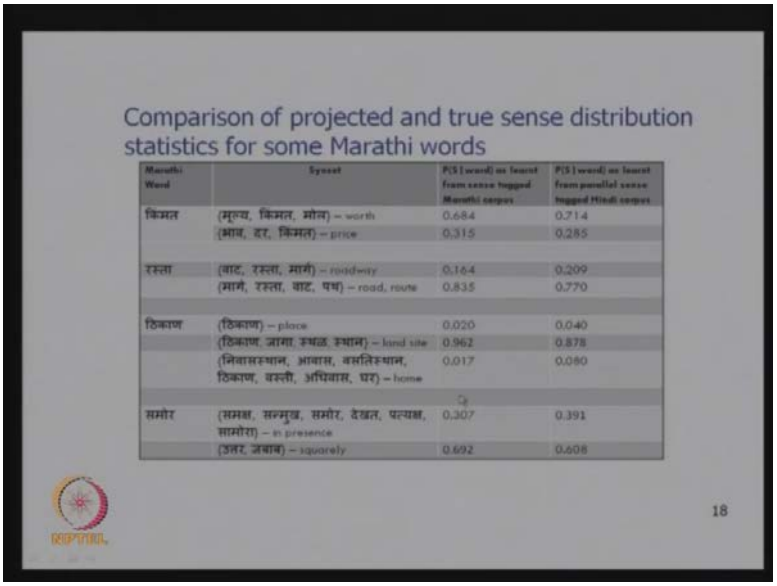
Let us look at some data, the Marathi word is kimat which has 2 senses worth and price, Marathi word is Rasta which has 2 senses road and route [fl] word is thikan, which has the sense of land site or place, and the sense is home also. Now, we actually had an sense of Marathi corpus and when, we calculate PS given the word P sense, given the word from Marathi corpus, the worth sense of [fl] comes with the probability of 0.684 whereas, the price sense comes with the probability of 0.315. If we use the Hindi sense of [fl] corpus for obtaining these values through projection and cross linking then we get the probability values as 0.714 and 0.285. I hope it is clear as to what, it is going on here the word [fl] is examine for its correspondence with the Hindi word, in the Hindi sense mart purpose, and the crossed linked Hindi word is obtained from the multidict.

So, these cross linked word appears, in the sense of word, how many times in the corpus, that is the question asked, and the cross linked to word in the sense of price, how many times does it appear, the cross linked word counts are found from the Hindi word corpus,

their ratios are taken, and we find the probability values. So, we find that the sense distribution actually obtained from Marathi corpus, is not very wide of the mark, compared to the sense distribution, in the Hindi corpus same is the case for [fl] a Marathi word [fl] the sense distribution is 0.164 for road way, and road or route sense has the sense of has the distribution probability value of 0.835. If we use projection from Hindi, than distribution is 0.209 and 0.770 similarly, for [fl] the home sense has a distribution of 0.037 and 0.962 for land site of place, when seen in the Hindi corpus, when these distribution is 0.878 and 0.12.

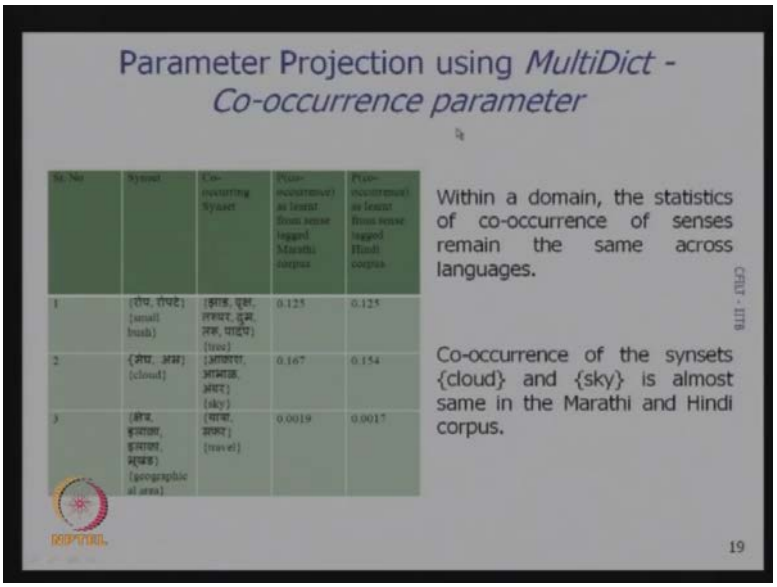
So, comparing the columns, we find that the values are wide up the mark, compared to the actual values we also did a calculation based on KL divergence and Spearman's correlation coefficient. And we found that the KL divergence, between the distribution learnt form a Marathi corpus, and the distribution learnt from a Hindi corpus comes out to be 0.2 9 which is not very large. So, in these the probability distribution from Marathi corpus, probability distribution from Hindi corpus, are not very distant from each other. Also the correlation coefficient is pretty high, between these values for Hindi Bengoli, the values are more striking. The KL divergence is only 0.05 between the probability distributions and Spearman's correlation coefficient is high, which is 0.82. So, there is a high degree of similarity, between the distributions, learn using projection and those learn from self corpus.

(Refer Slide Time: 28:23)



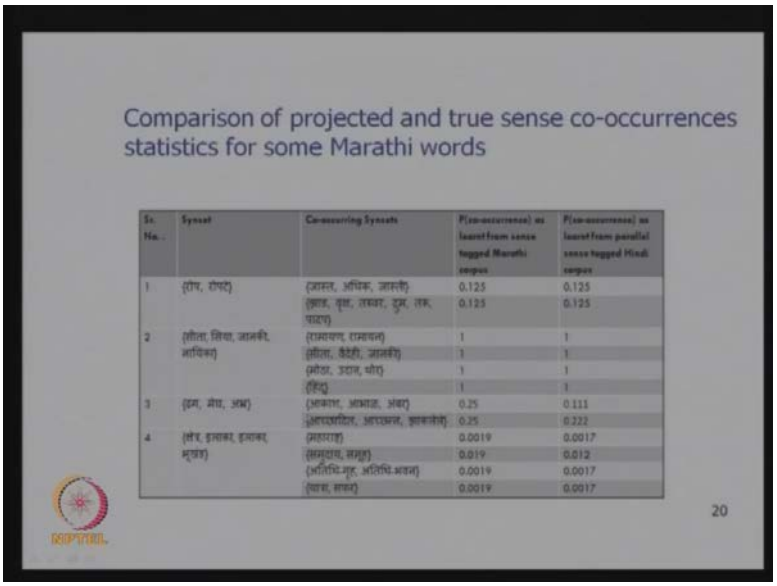
So, this shows the same statistics for a number of words.

(Refer Slide Time: 28:35)



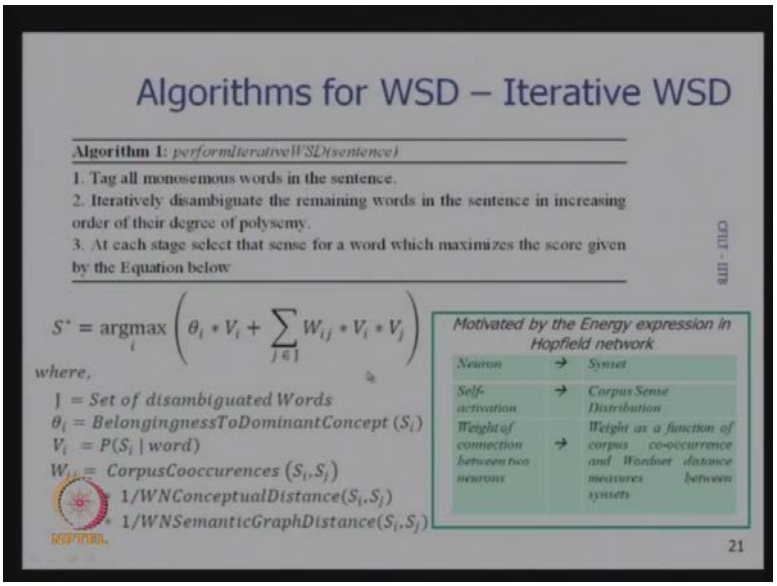
Now, we take the discussion to co-occurrence parameter projection, what co-occurrence co parameter is again a parameter, for words and disambiguation within a Domain. The statistics of co-occurrence of senses, remain the same across languages; this is our main foundational observation. Co-occurance of the synsets cloud and sky is almost same in the Marathi, and Hindi corpus few, we have few more examples; we take the synset rope, ropety which is a small bush the co-occurring synset is [fl] the sense of tree. Now the probability of co-occurrence is learn from sense Marathi corpus is 0.125 probability of co-occurrence is learn from the sense at Hindi corpus is 0.125 again [fl] which is the sense of cloud in Marathi has a co-occurrence of value of 0.167 with [fl] the sense of sky, from Marathi corpus, the corpus, the co-occurrence value comes out to be 0.167 when we see it in sense in that Hindi markov corpus it comes out to be 0.154. The final example is [fl] a geographical area, we examine its co-occurrence with travel, these co-occurrence value comes out to be 0.0019 when learn from Marathi corpus and 0.0017 when learn from Hindi corpus.

(Refer Slide Time: 30:27)



So, some more words are tabulated here, Sita Siya Janaki which is mythological character in the famous epic Ramayana, and it has a co-occurrence with, Ramayan and Sita matha Hindu and so on. And the distributions are identical weather they are learn from Marathi corpus or Hindi corpus. So, based on these 5 parameters, which are critical for words and disambiguation, we have come up with an algorithm for words and disambiguation. So, these algorithm, we call Iwsd the iterative words and disambiguation, and the algorithm makes use of some notions, Hopfield network in, this is shown in the slide here.

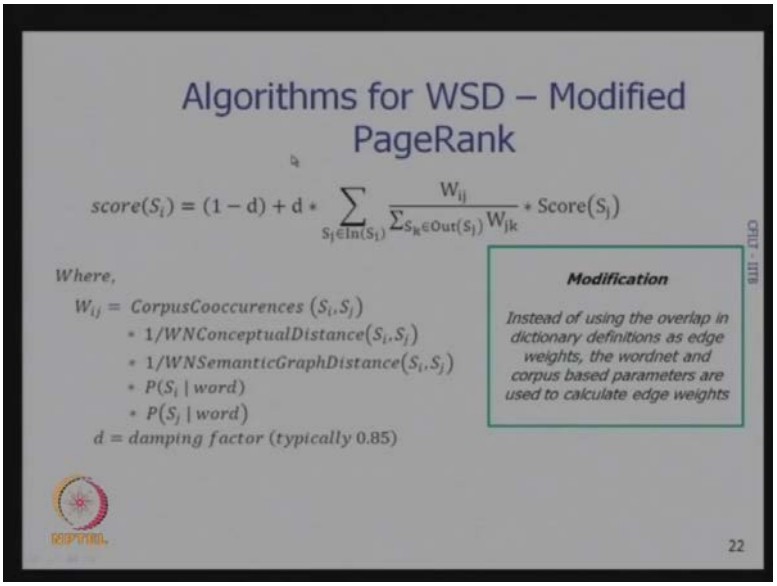
(Refer Slide Time: 31:36)



The scoring function us motivated by the energy expression in Hopfield network, the winning sense is obtained through a argmax competition by the scoring function. So, first we give the algorithm, we tag all the monosemous words in the sentence, we iteratively disambiguate, the remaining words in the sentence. In increasing order of their degree of polysemy at each stage, we select that sense for a word which maximizes the score, given by the equation below the scoring function is as follows V i's are the most important parameters Vi's are probability of Si given the word i s sense of the word, given the word J is the sense of disambiguated words, then theta i is a parameter, which captures the belongingness to a dominant concept S i, W i j captures the influence of the neighboring words.

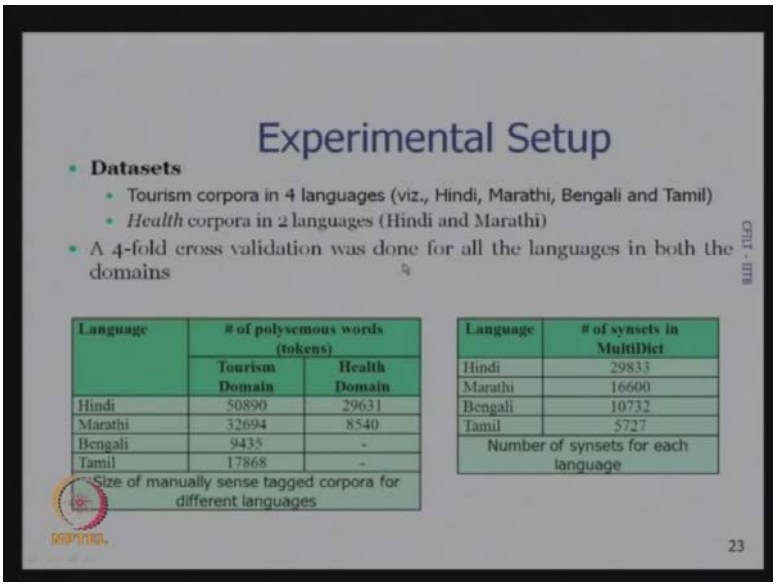
So, it is a product of 3 terms corpus co-occurrences, between S i and S j it is 1 by word net concept distance between S i and S j, and multiplied by 1 by wordnet semantic distance between S i and S j. So, clearly as the conceptual distance or semantic distance increase, the strong associations, association between the senses S i and S j reduces the strength of the association reduces. So, there is motivated by the energy expression of it Hopfield network. These are the correspondences the neuron corresponds to a synset self activation of the neuron corresponds to corpus sense distribution, weight of the connection between 2 neurons corresponds to weight as a function of corpus co-occurrence, and word net distance measures between synsets.

(Refer Slide Time: 33:50)



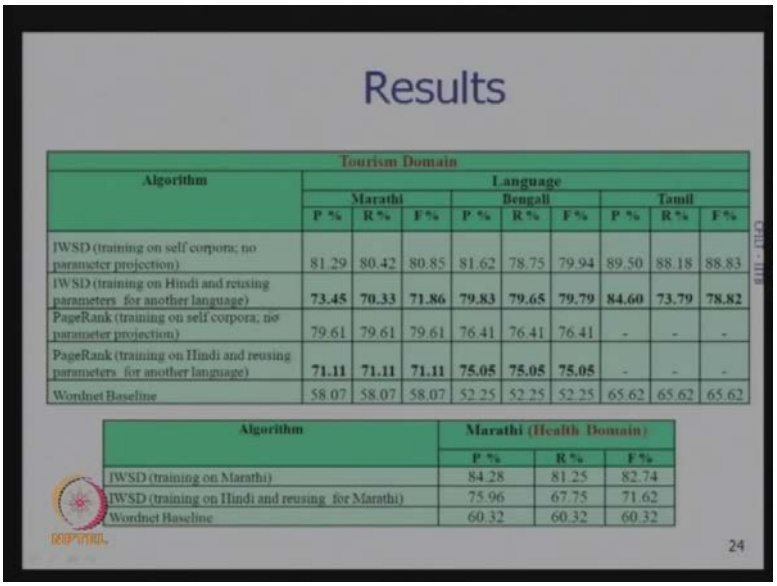
We have also used a modified Page rank algorithm as a scoring function, it has to compare the performance of or algorithm, with the page rank based percentage ambiguous the reason is that algorithms are very similar, we first disambiguate monosemous words, them disemous words and trisemous words and so on. So, the modification to the Page ranked based WSD as given by [fl] is as follows instead of using the overlap in dictionary definitions as edge weights, the wordnet and corpus based parameters are used to calculate edge weights. The Expressions here shows, the values for W i j which is measured in terms of corpus co-occurrence, wordnet conceptual distance, wordnet semantic distance, and the probability values of the senses given the words. So, these are cursive expression, just like in page rank calculation, score Si is expressed in terms of score S j, and there is a damping factor d.

(Refer Slide Time: 35:10)



The experimental setup was as follows, Tourism corpora in 4 languages were taken Hindi Marathi, Bengali and Tamil, Health corpora in 2 also were used Hindi and Marathi. A 4-fold cross validation was done for all the languages in both the domains. So, the first Language is Hindi, the number od polysemous words, is shown here in Tourism domain there are 50000 words. And in Health domain there are 29000 for Marathi in Tourism domain there are 32000 words and in Health domain there are 8000 words Bengali and Tamil only has words in Tourism Domain. And the statistics are as follows Bengali only 9000 and Tamil 17000, the number of synsets in multidict for Hindi are about 29000 Marathi is 16000 Bengali 10000 and Tamil 5000.

(Refer Slide Time: 36:16)

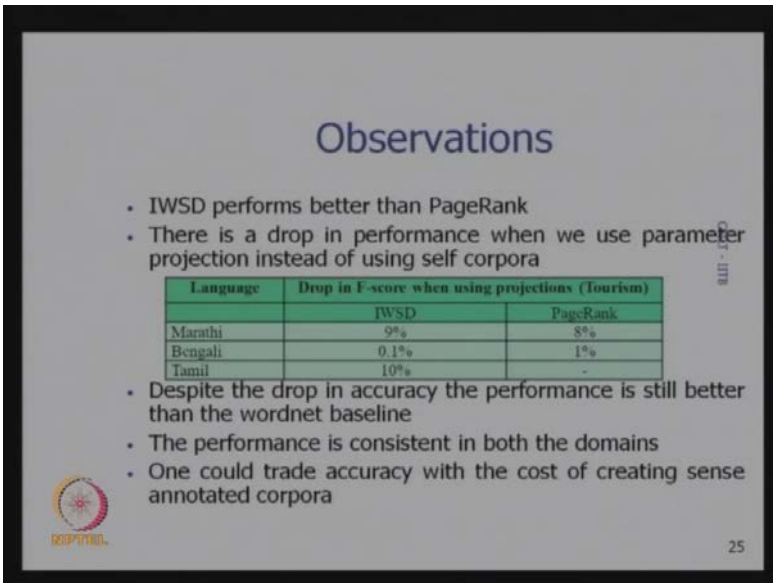


Now, we are ready to report the results of the algorithm with respect to various settings, first is the or IWSD algorithm the iterative w s the algorithm, which is trained on self purpose no parameter is projected and for Marathi. We find the precision is about 81 percent recall is 80 percent f score comes out to be close to 80 percent, more than 80 percent, let us concentrate on the Marathi for the moment. Now, trained on Marathi corpus the accuracy value comes out to be, the accuarcy value comes out to be, the accuracy value comes out to be 80 percent, when it is trained on Hindi and the parameters are used for Marathi. Then we find that the accuracy value comes out to be about 79 percent.

So, these shows that there is a fall in accuracy but the fall is not too much. Now if we take a similar algorithm the page rank algorithm which is trained on self corpora them, the accuracy value comes out to be about 79 percent. And if we use Hindi sense of corpora and reuse the parameters, than accuracy value comes out about 71 percent. So, what does it shows that, that when self corpora is not used, there is a fall in accuracy of about 10 points in both IWSD and Page rank algorithm. However the values are still far higher than the word net baseline that is the algorithm chooses, only the first sense of word net, independent of the context. So, there we find that the value is 58 percent, and there is a 20 point increase in accuracy on usage of self sense of corpus. And about 10 percent increase in accuracy, when Hindi sense corpus is used and parameters is projected.

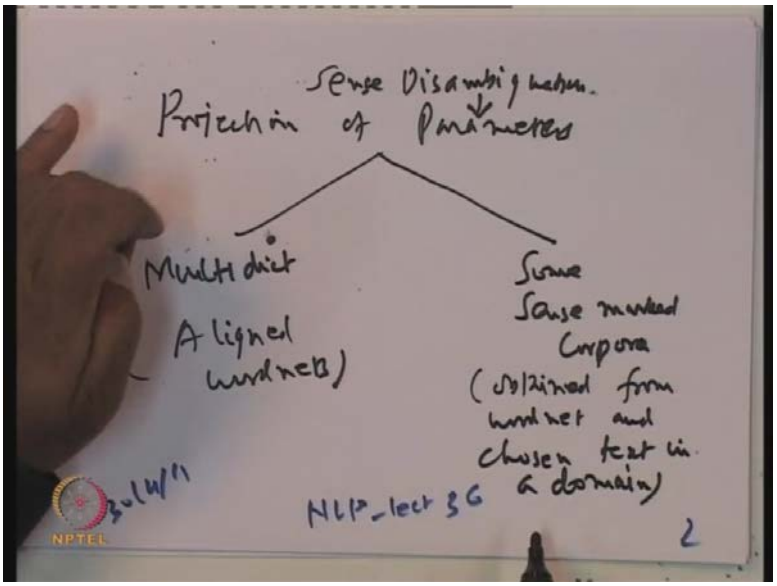
So, though there is fall in, about fall of about 10 percent in accuracy, the improvement is still very significant, the story is same for Bengali. In fact, here the parameter projection value hardly degrades for Tamil, there is a 10 percent degradation but the value is more than the baseline. When, we made our observations in the Health domain for Marathi, we found a 9 percent fall in accuracy, when senses when parameters are projected but there is a 10 percent improvement in accuracy, compare to baseline.

(Refer Slide Time: 39:43)



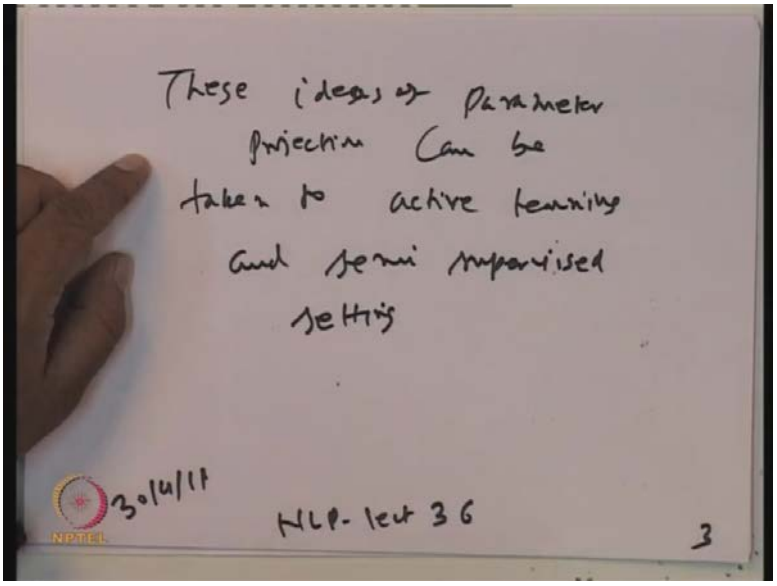
So, some observing, observations are IWSD the performs better than Page rank algorithm, there is a drop in performance when, we use parameter projection instead of using self corpora, for despite the drop in accuracy, the performance is still better than the word net baseline. The performance is consistent in both the domains namely dominance and namely Tourism and Health and one could trade accuracy with the cost of creating sense annotated corpus. So, this is an important point, which is made, being made here the point is that, we are making use of the work done for one Language for disambiguation of senses, in another Language. And these seems to have immense possibilities, Indian languages and many many languages in the world, like Turkish Hungarian Arabic and so on do not have too much of training corpus. So, if we can have sense mart corpora, of some of these languages, then they can be used for disambiguation of words, in other languages. The 2 things, which are needed for this are as follows, we will write it here, the first thing that is needed.

(Refer Slide Time: 41:27)



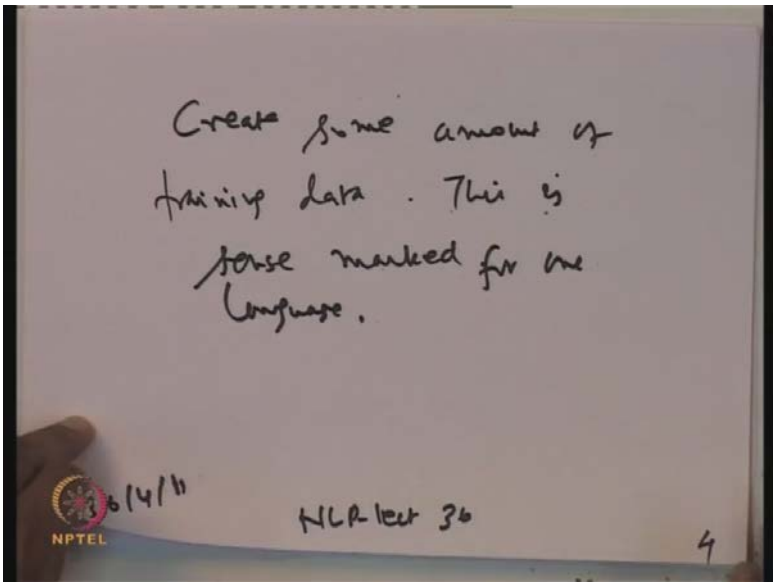
Projection of parameters, so sense disambiguation depends on projection of parameters what they require in term, is what is called the multidict, and some sense marked corpora, multidict is nothing but aligned wordnets sense marked corpora are [obtained] from word net, and chosen text in a Domain.

(Refer Slide Time: 42:45)



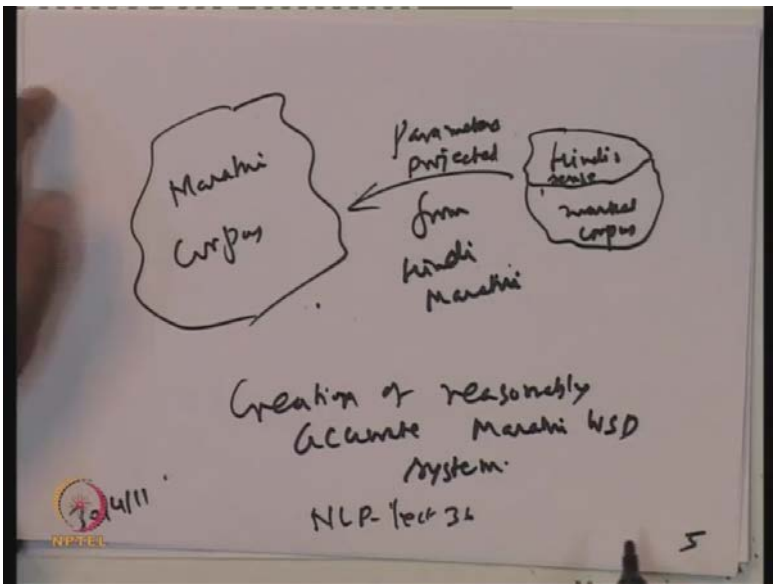
So, these ideas of parameter projection can be taken to active learning, and semi supervised settings, what we do is that.

(Refer Slide Time: 43:29)



Create some amount of training data, and this is sense marked for one Language, now comes the interesting possibility.

(Refer Slide Time: 44:12)



We have large amount of Marathi corpus, and we have a small amount of Hindi sense marked corpus. So, the parameters are projected from Hindi to Marathi, and that gives rise to creation of reasonably accurate, Marathi WSD system. So, these ideas are fairly new, and have been explode by us in detail, and reported in top level research flora so with this, we finish our discussion on words and disambiguation. We would like to

summarize by saying that word are based algorithm, supervised algorithm, semi supervised algorithm, unsupervised algorithm, these approaches have been discussed in detail. And finally, we have taken a look at, the new direction of words and disambiguation research, which is in presence of small amount of training corpus or no training corpus. However, a very strong requirement is that, we should have aligned wordnet for different languages, and though we say that the languages resource constrained, we still have to invest creation of which means, we still have to invest in creation of word net, which is by no means a small task.

Now, an interesting question that arises is that what does words and disambiguation research read towards, is it really crucial, words and disambiguation is important for many applications like question answering, than accurate machine translation them summarization and so on. But there is a view, in natural Language processing community that if we have lots of training data, and capture many different contexts, than words. And disambiguation may not be required or may be required in a minimal amount. This is not really critical, if we have large amount of training corpus and sophisticated machine learning algorithm. But this is a debate which going on in the natural languages processing community, and only future developments will show where the truth lies, next will discuss parsing.