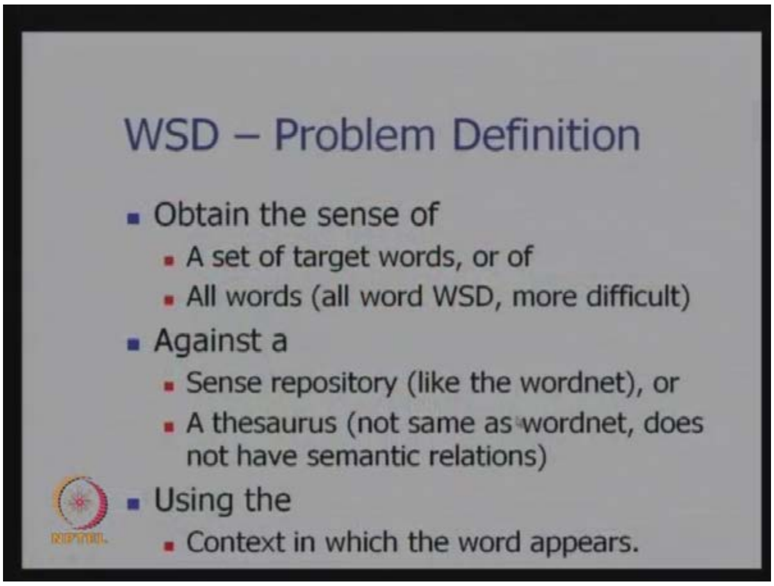
Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 33
Word Sense Disambiguation; Overlap Based Method; Supervised Method

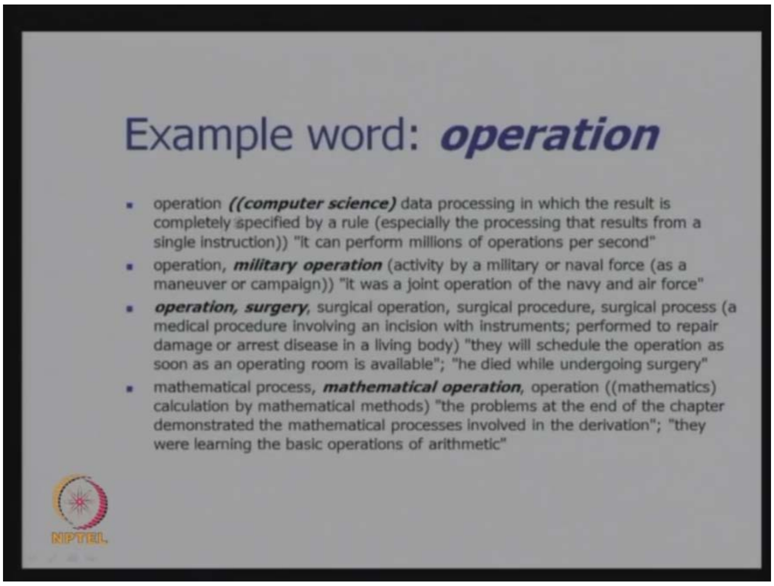
We will continue discussing the word sense disambiguation methods. And today's focus will be on discussing overlap based method in more detail, and also start the supervised method.

(Refer Slide Time: 00:36)



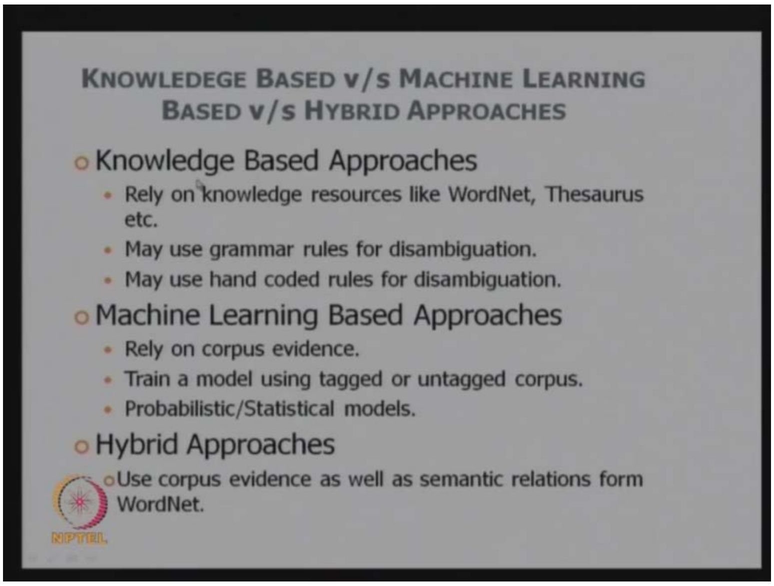
So as we have seen before word sense disambiguation is the problem of obtaining the sense of a set of target words or of all words, which is called all word WSD is a more difficult problem. Against the sense repository, like the wordnet or a thesaurus, not same as wordnet, does not have semantic relations, the thesaurus does not have semantic relations. Using the context in which the word appears.

(Refer Slide Time: 01:14)



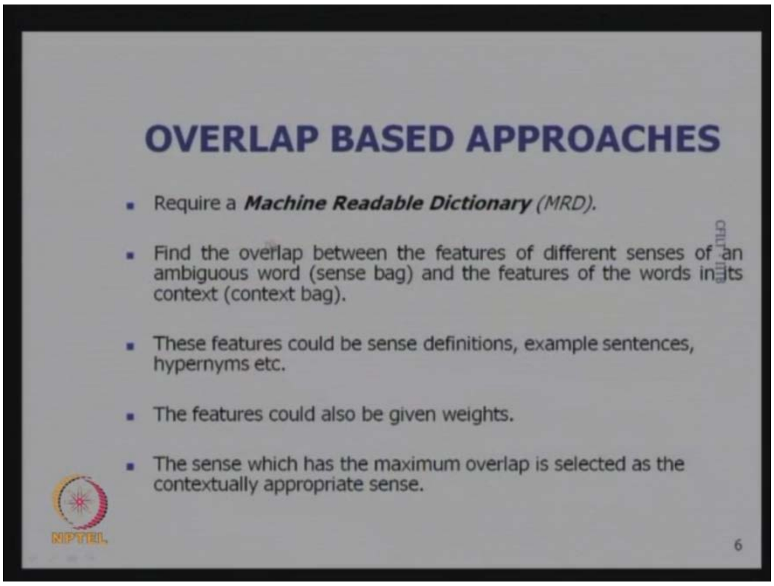
So, the example, which is running is the example of the word operation, which can have a computer science sense or military operation sense or medical operation sense or mathematical operation sense. So, the goal will be to find out the sense of the word as detected by the context.

(Refer Slide Time: 01:36)



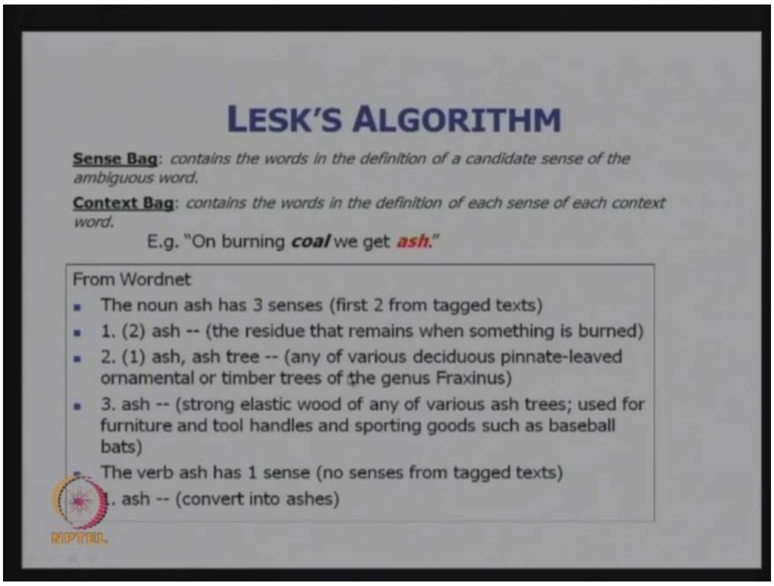
Now, we have been discussing the knowledge based methods, machine learning based approaches will be covered after this, and hybrid approaches. In the knowledge based approach, we rely on the knowledge recourse like, the wordnet thesaurus etcetera. We may use the grammatical rule for disambiguation, we may also use hand coded rules for disambiguation. Then in the machine learning based approaches, we need essence part of corpora, from which, we have to produce the corrections, and in hybrid approaches, both corpus and semantic relations are considered. So, knowledge based approaches is covered, and overlap based approaches is one kind of such a approach. Now, in overlap based approache.

(Refer Slide Time: 02:22)



As the slide shows, we require a machine readable dictionary, we find the overlap between the features, of the different senses of an ambiguous word, the sense beg, and the features of the words in the context, we called it the context bag. These features could be sense definitions, example sentences hyponymy etcetera, and the features could also be given weights, the sense which has the maximum overlap is selected as the contextually appropriate sense.

(Refer Slide Time: 02:52)

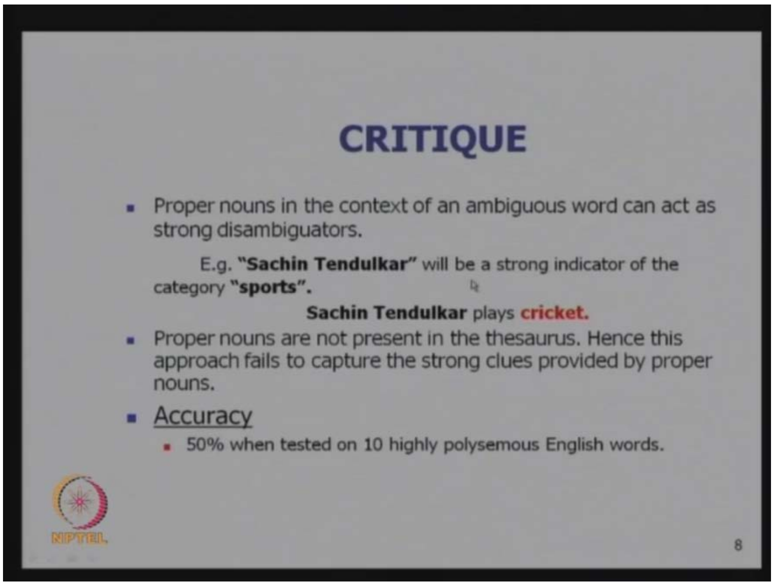


Now, an example was discussed in last time, we are repeating this, now the sentence which is given is on burning coal, we get ash. So, here the, what we disambiguated it is ash, this is shown in red, the contextual clues, are the word power and coal. Now, sense bag contains of words in the definition of a candidate sense, of the ambiguous word. And it the context bag contains the words in the definition of each sense, of each context word, so here we are going to disambiguate ash, so from the wordnet, we found find that, noun ash has 3 senses. The first sense is the most commonly known sense of ash, the residua that remains, when something is burned, so this is the first sense.

Second sense is that of ash tree, and third sense is that of elastic wood, so we have to find out, from the context, which sense applies in this particular case. The first sense is the residue sense, obtain after burning, second sense is the tree sense, third sense is the wood sense. So, when we read the sentence once again, on burning coal, we get ash, we find that, it is the first sense which is the applicable. Now, from the contexts we find that burn and coal and get, are the contain words, in the environment, and in the senses bag, for each sense, we take a gloss, which is reference of senses, and the example sentence. And find out if the context bag has overlap any of them, and which one has the highest overlap.

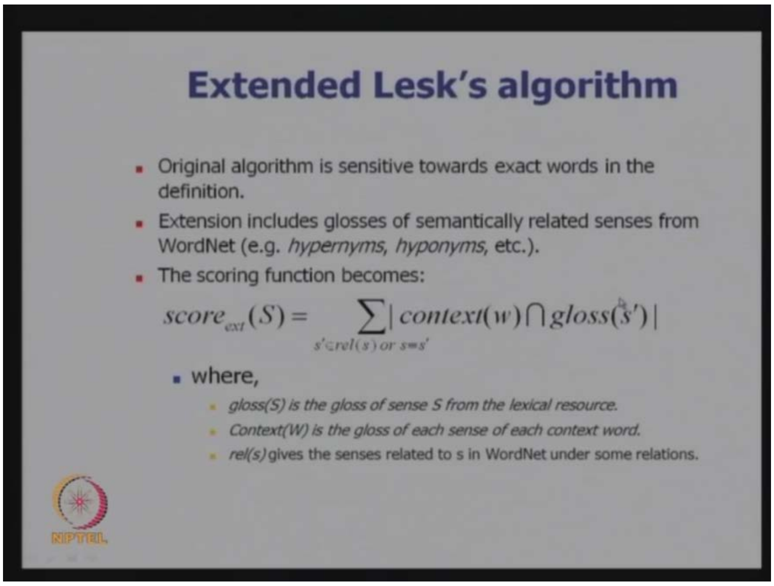
So, if you look at the first sense of ash, the gloss is written as residue remains, when something is burned. So, the context words, which are burned coal and get, have found and overlap in the definition, the first sense has the overlapping, in the formal path. No other definition, as you can see here, has a overlap, so we declare the first sense has the winner, and we will be correct here. So, the first sense is the winner, and this is the sense, in which the word ashes used.

(Refer Slide Time: 05:28)



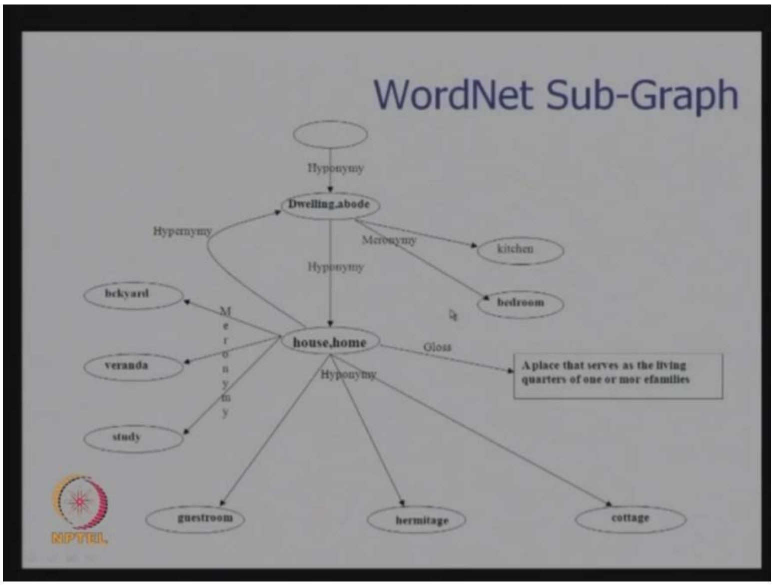
So, this was discussed, and we have seen that, the main critique is that proper nouns are offend strong indicators, of the senses, and they are not used in this method of disambiguation. So, the other point is that, there can be Sevier drop topic drift, because of wrong overlap.

(Refer Slide Time: 05:55)



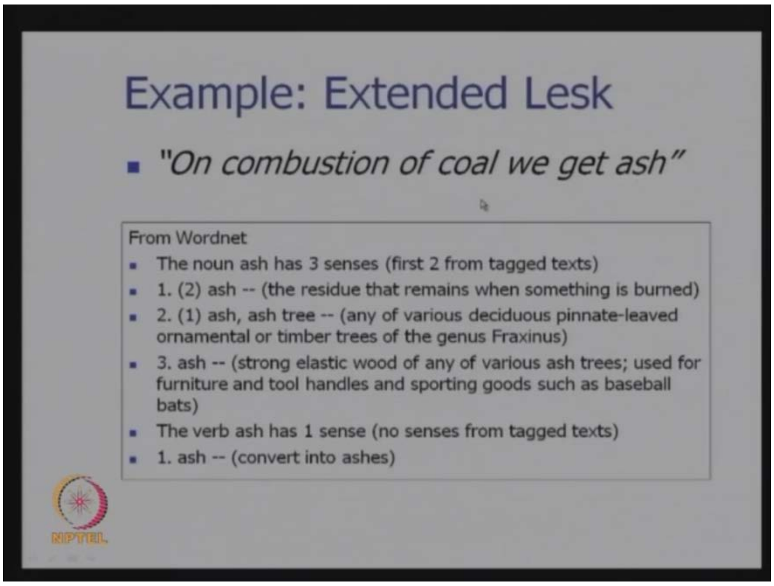
Then we may use of, then we studied the Extend Lesk's algorithm, where the overlapping is found, not only from the words on seen sense, and the gloss and example sentence.

(Refer Slide Time: 06:12)



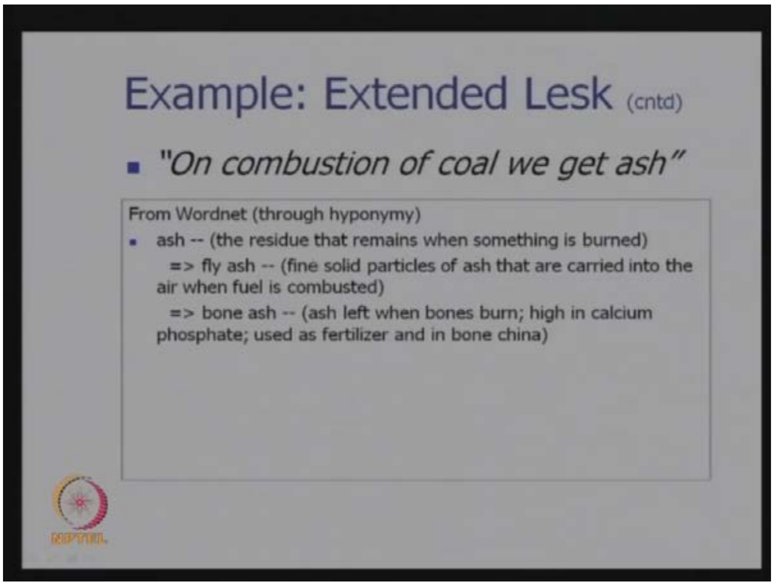
But also from the related semantically related notes, so one would like to see the overlap, from Hypo Hyperymy, Hyponymy, Meronymy, colonymy and so on. Wherever, there is connected note, we would like to go in that note, obtain the words in that note, create the senses bag, and then find out the overlap. So, this is the excellent work Lesk's algorithm, and here the main problem is that, the clue words, which come from the related notes, can again cause tropic drift, and mislead the algorithm, to believe that this is the sense.

(Refer Slide Time: 07:00)



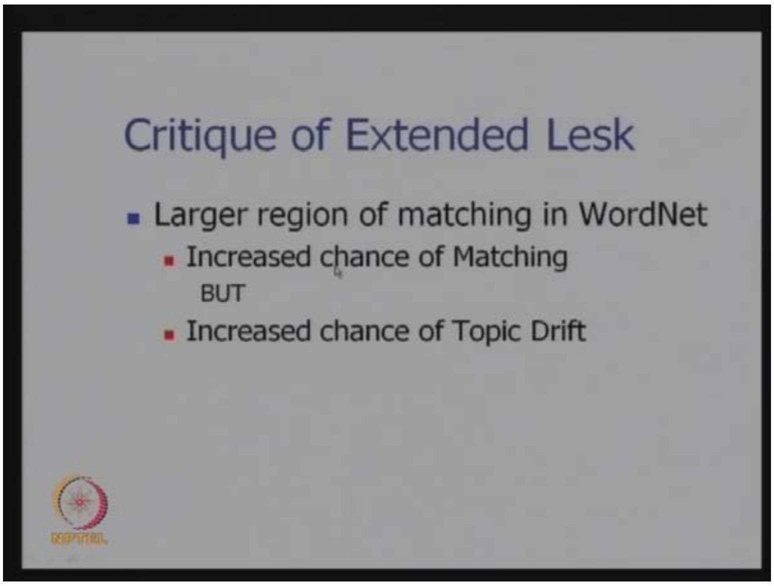
So, and example was shown in Extended Lesk on combustion of coal, we get ash, so in its own note, there is no overlap, because we do not have the word burning here, even though it is synonymous.

(Refer Slide Time: 07:20)



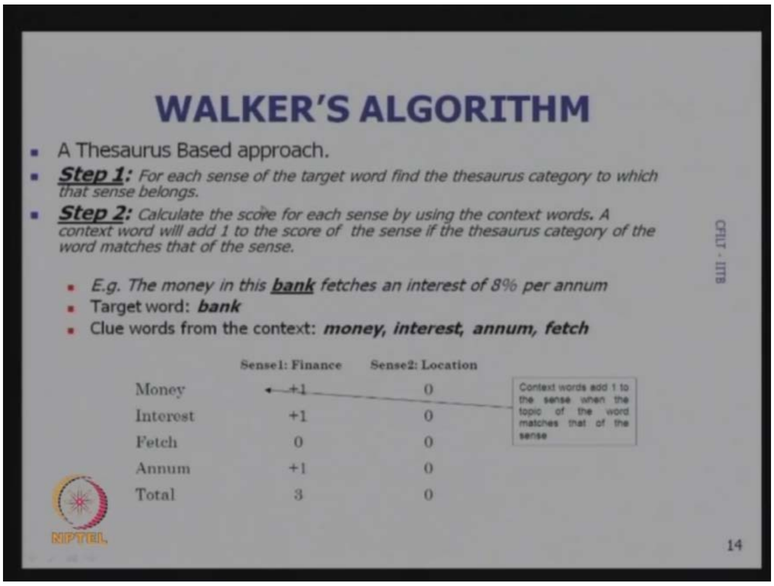
But we do get, a get at overlap when we follow the hyponyms semantically relationship. So, residue that remains, when something is burned, and fly as the kind of ash, where we find the definition is fine solid particles of ash, that are carried into the air, when fuel is combusted. So, combusted has an overlap with combustion, so this is the winner sense.

(Refer Slide Time: 07:49)



So, these things were discussed, and the critique is that, there can be topic drift.

(Refer Slide Time: 07:54)



Now, we will move on to another algorithm, which is known as the Walker's Algorithm and we discuss it in detail.

So, it is important to understand that, Walker's Algorithm and subsequent algorithm, which is conceptual based algorithm, yet another algorithm, which is Page Rank based algorithm, these are different interesting ideas, algorithmic ideas, which are applied to weight the sense. So, these are mainly weighting mechanisms or scoring mechanisms, to bring to the 4 important particular senses, for purpose of disambiguation. But it is important to understand, that fundamentally at the basic level, these are nothing but overlap based algorithm. They essentially measure the overlap but give a principle, principle way of a scoring a particular sense.

So, let us go ahead, with the Walker's Algorithm, this is a thesaurus based approach. So, this is not really based on wordnet, wordnet may assist this algorithm, which is an interesting idea but it is based on the thesaurus.The first step is that, for each sense of the target word, find the thesaurus category, to which the sense belongs, And step 2 is calculate the score of each sense, by using the context words. A context word will add 1, to the score of the sense, if the thesaurus category of the word matches, that of the sense. For example, let us take the sentence, the money in the bank, fetches an interest of 8 percent per annum, the target word here is bank, which we know has 2 senses one is the riverbank sense, and other is the financial bank sense.

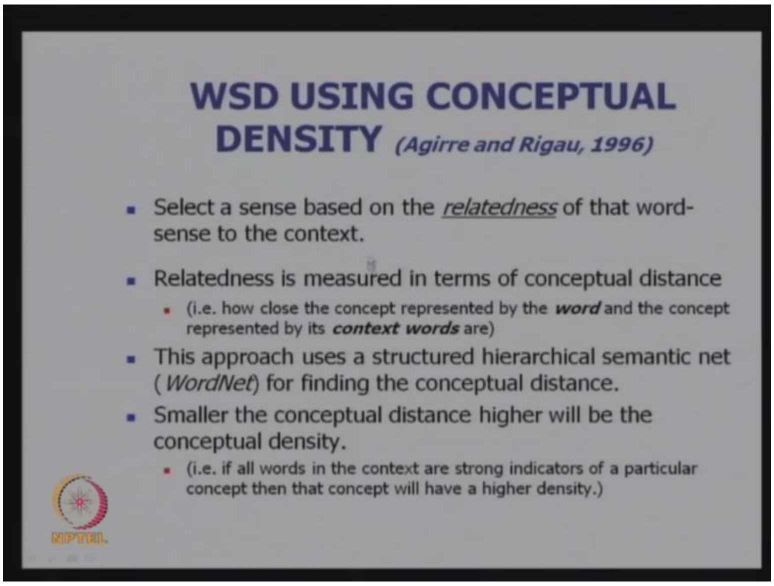
Now, what does the common senses tell us, we see that, in this sentence, there are words which relate to financial domain, money for example, is the word which does that, fetches and interest, interest is again a anonymous word, and one of its sense is, in the financial domain, 8 percent per annum. So, per annum is also a fixed fetch, which is highly used, in the financial domain 8 percent is rate of interest. So, the clue words here, are money interest annum at fetch, fetch is also of course, an important clue, because this is a involve a content word, in the context. So, here is how the algorithm goes, we open as many counters as there are further senses of the target word. Now, let us assumes here, that the bank has already the known bank has only 2senses, that of the financial sense, and the location sense. So, what is happening here is the moment, we see money, we add the count of finance counter by 1, we increment the finance counter by 1, and money has no sense of locations, so nothing getting gets added to sense 2.

Interest similarly, has many senses but one of the senses is in the financial domain that increments the count by yet 1 more 1. I am nothings gets at a 2 location sense, fetch does nothing to either sense, annum is a financial term, and so this adds 1, again to financial counter, nothing gets added to the location. So, after we have made use of all the annum does also does the same thing, after we have made use of all the clues, we find that finance counter, value has become 3 location counter value is 0. This is an allegiants simple idea, which says that, for each sense gone increment the counter, from the overlap up, the context word, and thereby find out the winner.

So, the problem here, though is that, though the algorithm is elegant, the problem is who will say, that money fetch annum, they have finance domain sense, how do you know that, is there a data structure or a knowledge based, which will facilitate that. The answer to the question is yes, there us something called an anthology, which is independent of language, and which is purely devoted to organizing, the meaning targets in a situation to a hierarchy. So, this anthology is anthology of flexible financial domain, and the anthological notes, which are the properties of different senses, regional property of different senses.

They will denote that annum money interest fetch, they are financially relevant words, they belong to the domain annum finance, and we have to make use of that anthological structure, the thesaurus itself mentions in the senses of the words, the anthological information. So, from this information, it is possible to get the clues of the words, in terms of their domain dependence, and we can go on incrementing the counter, as described in the algorithm. So, this is the main idea of the Walker's Algorithm. So, we look at the slide here, then see that the finance has counter value increment into 3, location counters is 0.

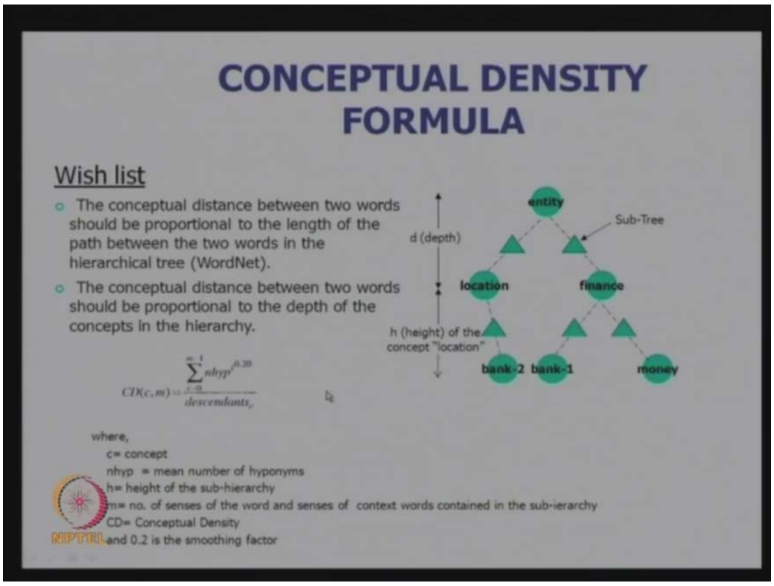
(Refer Slide Time: 14:21)



We move on to, the next algorithm which is the conceptual Density Based Algorithm proposed, by Agirre and Rugau back in 1996. This algorithm again has, a few interesting ideas, and the main idea is that, we have to select a sense, based on the relatedness, of that word sense to the context. So, these a fundamental idea, and relatedness, how was is measure, it was measure in overlap based approaches, in terms of the overlap between the sense bag and the context bag.

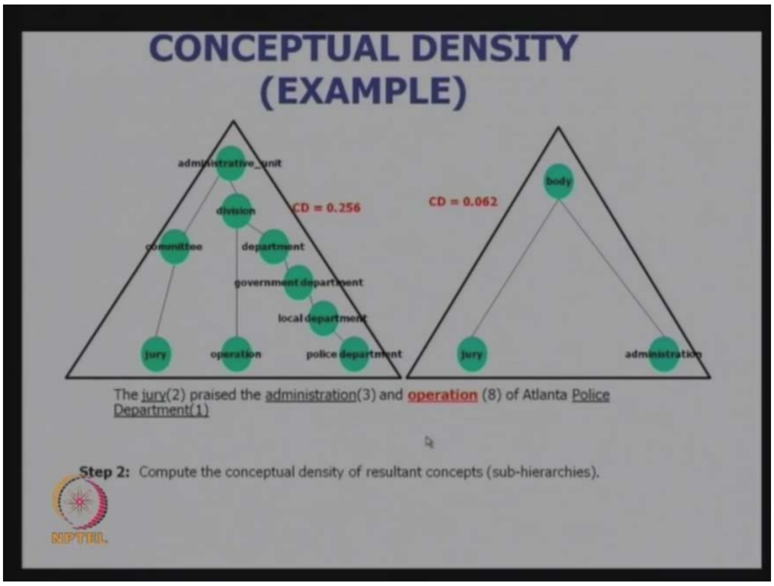
Now, in case of conceptual distance or density on the relatedness is measured, in terms of the conceptual distance, that is how close is the concept, represented by the word, and the concept represented by its context words, are the approach uses a structured hierarchical semantic net, in this case its a wordnet for finding the conceptual distance. Smaller the conceptual distance higher will be the conceptual density, that is if all the words, in the context a strong indicators of a particular concept, then that concept will have a higher density, so will illustrate this.

(Refer Slide Time: 15:40)



By means of an example, we will come to this particular slide, after sometime.

(Refer Slide Time: 15:51)

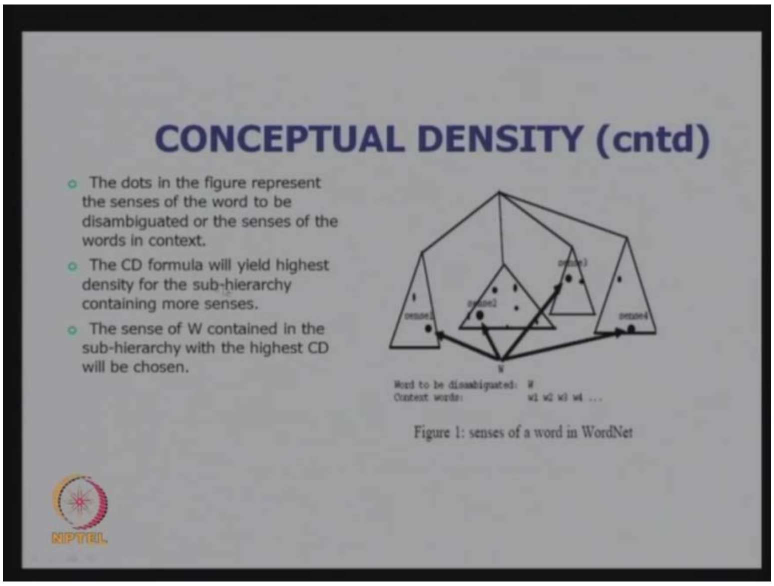


We have taken here, an example sentence, which is the jury press, the administration and the operation of Atlanta police department, this example is from the paper by Agirre and Rigau. So, here we like to go back, to the wish list the, is that, the conceptual distance between 2 words, should be proportional to the length of the path, between the 2 words in the hierarchical tree. And here the tree is wordnet, the conceptual distance between 2 words, should be proportional to the depth of the concepts, in the hierarchy. So, if the 2 notes are high up in the hierarchy, then their conceptual distance also, should be more. So, because of this consideration, we come up, with the scoring mechanism, which is for the conceptual distance of concept, with respect to m senses, in the context of the word.

So, this is the formula, the formula is CD c comma m is equal to sigma i equal 0 m, m equal to 0, i equal to 0 to m minus 1, and the n hype to the power i to the power 0.2.0 divided by descendents of c, c is the concept the n hype is the mean number of hyponemas, h is the height of the severity, m is the number of senses of the word and senses of the context word, contain in the severity, c d is the conceptual distance and 0.2 is the smoothing factors.

So, this formula is used for finding the correct sense, with respect to conceptual density, that means higher the score, of the particular concept, more is the chance, this will emerge as a winner sense. So, here is an example of the situation involve, n t t is the route of everything and under that terebreiya sub tress, you can see here, in the financial domain, this bank one is the first sense of bank, in the financial domain, there is this what money also. Then the second sense of bank is the riverbank sense, which is which comes under location, and it is shown here.

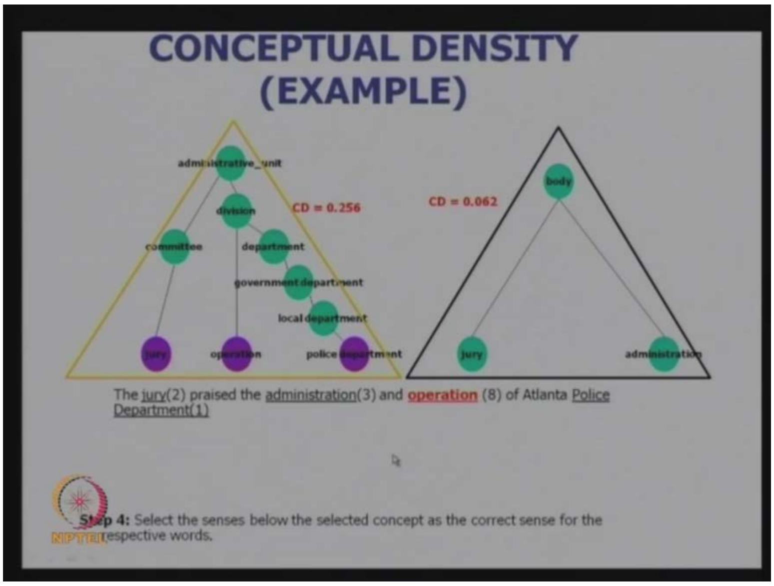
(Refer Slide Time: 18:44)



The conceptual based density, words sense disambiguation sense algorithm works as follows, the dots in the figure represent the senses of the word, to be disambiguated or senses of the words in the context. So, first one is done is that, the senses of all the words, both target, and non target are collected, this happens by means of a window, around the target word. Now, which collect the words, we build the sub tresses from the wordnet hierarchy, and whichever, sub tree has maximum conceptual density, as per the formula given before, becomes the winner sense.

So, those are the senses, which will be picked up, from the window, as the winner senses. So, this is an all the word senses, disambiguated algorithm. So, that this algorithm happens for all the words, for all possible senses, and what we do is that, we note that number of notes, in a particular sub tree but we also give weighted to how high the concept, is in the conceptual hierarchy. And, we also give weighted to the descendents, and so on. So, based on this we finally obtain the correct senses of the word.

(Refer Slide Time: 20:27)

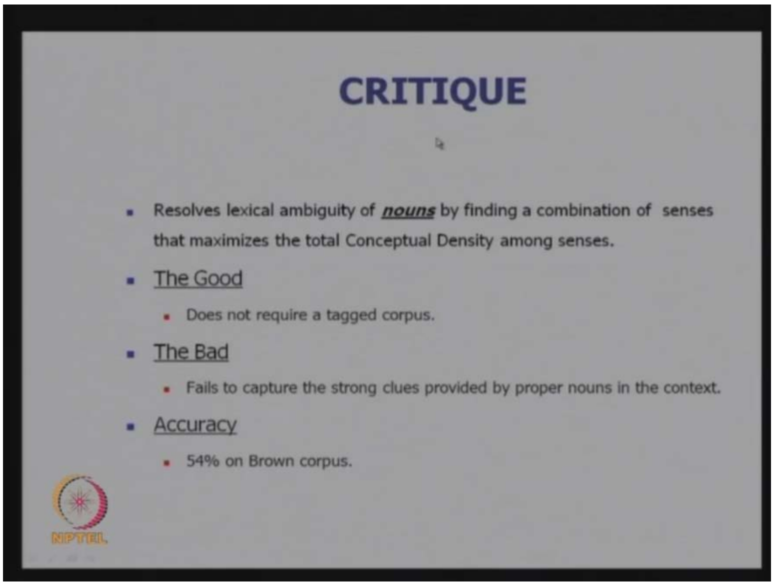


So, here is an example, the sentence is the jury praised the administration, and operation of Atlanta Police department. So, here the words sense disambiguated are jury, administration operation and police department. So, jury has 2 senses, administration has 3 senses, operation is ambiguous very ambiguous this has 8 senses, and police department is 1 sense. So, first we make a lattice of the nouns in the context, and their senses at hyponyms. So, we can see here that, administrative jury, the word jury has hyponymy as committee, which again has hyponymy, as administration unit, this jury has hyponymy as body. So, we are taking about 2 senses of the word jury, then we find out the sentence, other entity also in the context of the sentence.

So, administration is related to body, operation is related to division devisor again is related to administrative unit. After that Police Department is taken, which has local department has hyponymy, which has government department has hyponymy, then department has hyponymy, then division then administrative unit. So, this way, we have build a conceptual sub tree, which is routed as administrative unit, and another sub tree has been built, which us routed as body. So, when this 2, this 2 trees are made, we are ready to compute the conceptual density, of the resulted concept or the sub hierarchies, and the concept which the, with the highest CD is selected.

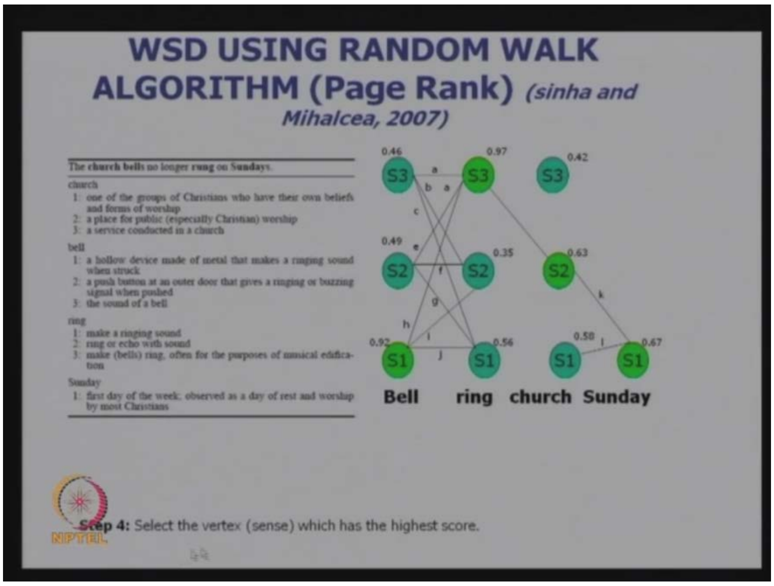
So, this is the sub tree, which has highest conceptual density, and the senses will be picked up from here. So, a police department has a single sense, we do not have to worry about this, administration has 3 senses, and we can pick up the sense from there, and jury has 2 senses but this jury here, has much more dens sub tree, compare to this jury here, so this will be the winner sense. So, this is the main idea of Conceptual Density based algorithm, for senses disambiguation, here the conceptual density is 0.256 at here the conceptual density 0. 062 as per the formula discussed before. Select the senses below the selected concept, as the corrections for the respective word.

(Refer Slide Time: 23:37)



Now, the critique of this approach is that, it resolves lexical ambiguity of nouns, by finding a combination of sense that maximizes the total Conceptual Density among senses. The good part of this algorithm is that, it does not required a tagged corpus, which is difficult to create, the bad part is that it fails to capture the strong clues, provided by proper nouns in the context. So, this is the perennial problem, for all overlap based approaches, and the accuracy of this approach is that, it is 54 percentage on the brown corpus, the brown corpus is, the one which gives us gives right the same corp, the sense corpus crafted out of the brown corpus, and this is this has been a bench mark data, for many WSD algorithms. So, proceeding further.

(Refer Slide Time: 24:33)



We now take another algorithm, which is Word Sense Disambiguation, using Random Walk Algorithm, and this is based on the idea of Page Rank the Algorithm was propose in 2007 by sinha and Mihalcea. We take this example here, of the sentence shown in the screen, the church bells no longer run on Sundays, church has a 3 senses, the first sense is so one of the group of christian, who have their own believes, and forms of worship. The next sense is a place for public, especially christian worship, and 3 is service conducted in a church.

So, the church bells no longer rang in sunday’s, most likely, it is the second sense, which is applicable here. There is a bit of uncertainty, the bit of uncertainty, with respect to 2 and 3, it could be could it be the service sense but the most likely you know, it is the place sense, which is used here. Bell also has 3 senses a hollow device, made of a metal that makes a ringing sound, when struck, a push button at an outer door, that gives a ringing or bugging signal, when pushed the sound of the bell. So, the sense which have applicable here, is the first sense a hollow device. Ring also is 3 senses making a ringing sound, ring or echo with sound make bells ring often for the propose for musical ratification. So, here the first sense in applicable, make ring sound, sunday is the word another contain for, this is the first day of the week, observed as a day of fresh and worship by most christian, sunday does not have too many senses just 1sense, so it is bonus emus.

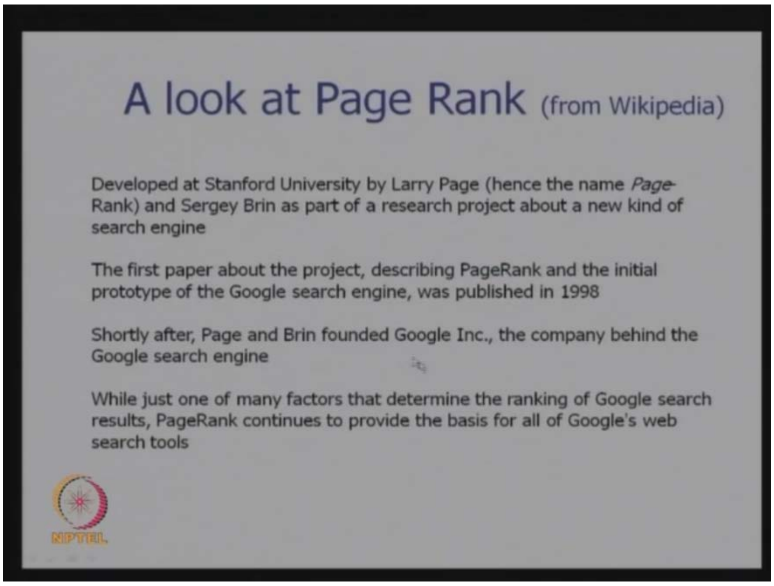
Now, our goal is to see out of these 3 into 3 into 1 which are 9 senses, which senses are applicable, that know what but cannot the algorithm find it out, for church it is sex sense, 6 second sense for bell, it is first sense ring also it is the first sense. So, here is the graph which gets spelt slowly, for a each word, in the sentence where ring church in Sunday, which erect a column of senses on the words. So, bell has 3 s1 s2 s3 ring as 3 senses s1 s2 s3, church has 3 sense s1 s2 s3, and sunday has a single sense.

So, add a vertex for each possible sense of the each word, in the context, next add weighted ages, using the definition ways semantic, similarity Lesk's method. So, which add weights from 1 word to the next word, to the senses columns . So, here we build the edges from s3 to s3 of the next ward, s3 to s2 next ward, s3 to s1 of the next ward, similarly, from s2 and s1. So, we will have 3 to 39 arches, growing from bell to bell's senses to ring's senses. So, this the weighted of this edges, whole be equal to the overlap based, similarity between the senses s3 or s2.

So, we will take the senses, seen s2 so and they take the gloss words, and take the example words, and then see how much is the overlap, between the corresponding entities, in the order sense, and record that as the weighted of the arch, we do this for all that 9 arches. In the step 3, we used the graph based tracking algorithm, to find the score of each vertex that is for each word sense. So, when this algorithm is run, then we find that, at that conversions these are the winner senses.

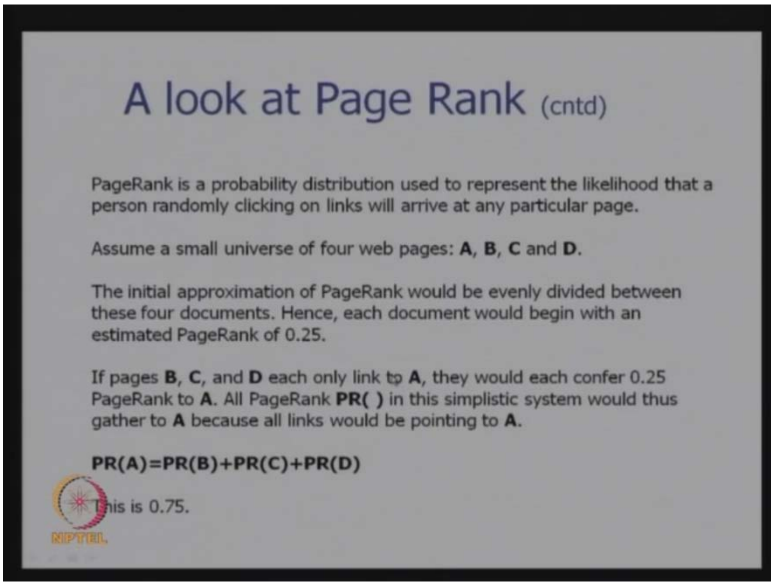
So, s1 is the winner sense for bell s3 is the winner sense, for ring make bells ring, often for the purpose musical ratification. We stand corrected here, we earlier said it is the first sense, which was applicable, in this particular context, it is the third sense, which is applicable, and for church it is the second sense, which is applicable in context, similarly, for Sunday there is only 1 sense which is s1. So, these are the senses, which have become winner, after the Page Rank Algorithm has been, so select the vertex which has the highest score.

(Refer Slide Time: 30:02)



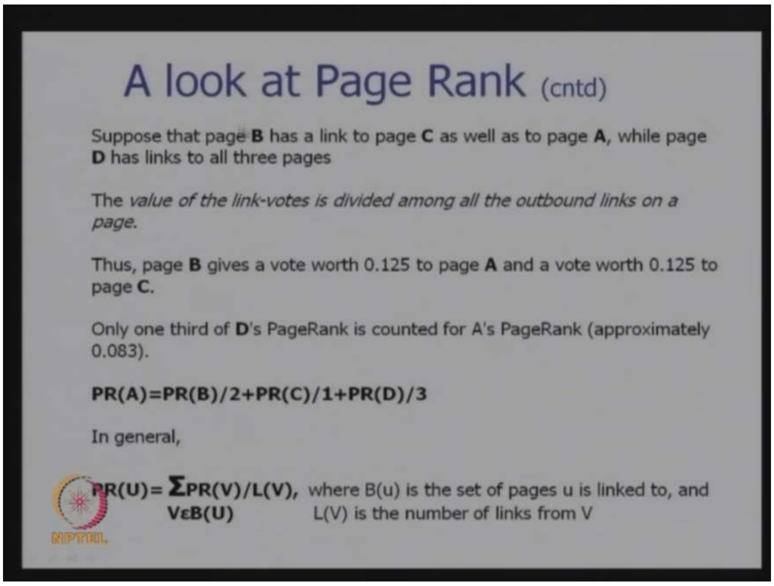
Now, this is the main idea but what happens actually, at the algorithm level, we look at the Page Rank idea, and we make those of the Wikipedia description. Page rank is Developed at Stanford University by Larry page, hence the name Page rank and Sergey Brin as part of a research project about a new kind of search engine, the first paper about the project describing, Page Rank, and initial prototype of the Google search engine, was published in 1998. Shortly after Page and Brin founded Google incorporate, the company behind the Google search engine, and while just one of many factors, that determine the ranking of Google search results, Page Rank continues to provide, the basis for all of Google’s web search tools, so this is a bit of historical remark on Page Rank.

(Refer Slide Time: 31:02)



What actually happens in the algorithm, it is page rank is the probability distribution, used to represent the likelihood that a person, randomly clicking on links, will arrives at any Page, Page Rank try to capture, this particular intuition. It is probability distribution used to represent the likelihood, that a page rank person randomly clicking on links, will arrive at any particular page, for illustration let us assume, a small universe of 4 web pages A B C and D. The initial approximation of Page Rank, would be evenly divided between, these 4 documents. Hence, each document, would begin with an estimated Page rank of 0.25, so if pages BC and D each only link to A. They would each confer, 0.25 Page Rank to A, all Page Rank PR in this simplistic system, would does gather to A, because all in links would be pointing to A. So, Page Rank of A, is Page Rank of B plus Page Rank of c plus Page Rank of D, which comes out to be 0.75.

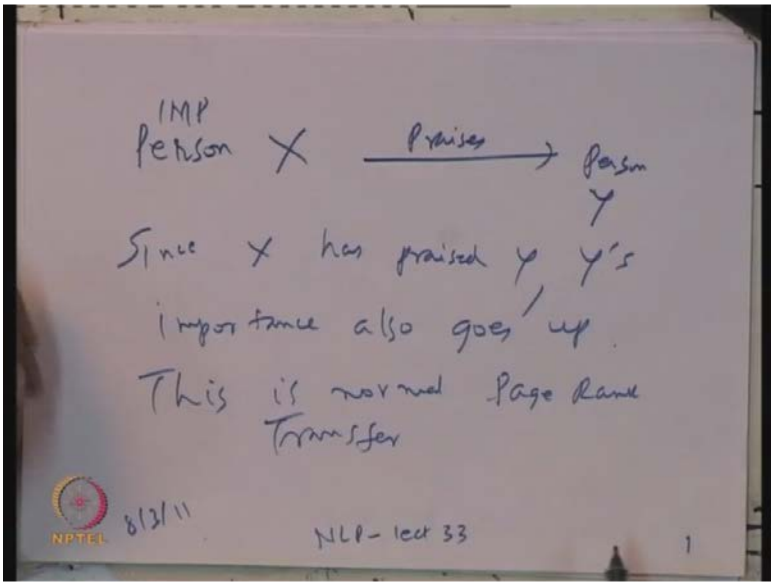
(Refer Slide Time: 32:29)



Now, suppose that page B has a link to page C, as well as to page A, while page D has links to all 3 pages. So, the value of the link votes is divided among all the outbound links on a page. Since B has link to page C as well as to page A. We divide D's page rank into 2 parts, and give 0.125 to A, and 0.125 to C, dividing 0.125 into 2 equal parts. Similarly, since D links both A and B and C, A D's page rank is also divided into equals 3 parts, and given to A. So, the modified formula for page rank of A, compare to the last formula is this, Page Rank of A is equal to Page Rank of B divided by 2, Page rank of C divided by 1, and Page Rank of D divided by 3. So, this is the modified formula, earlier you remember there were no dividing factors. In general Page Rank of note u is equal to Page Rank of v divided by l v, summing over all possible v is belonging to b u, where b u is the set of pages, u is link to, and l v is the number of links from B.

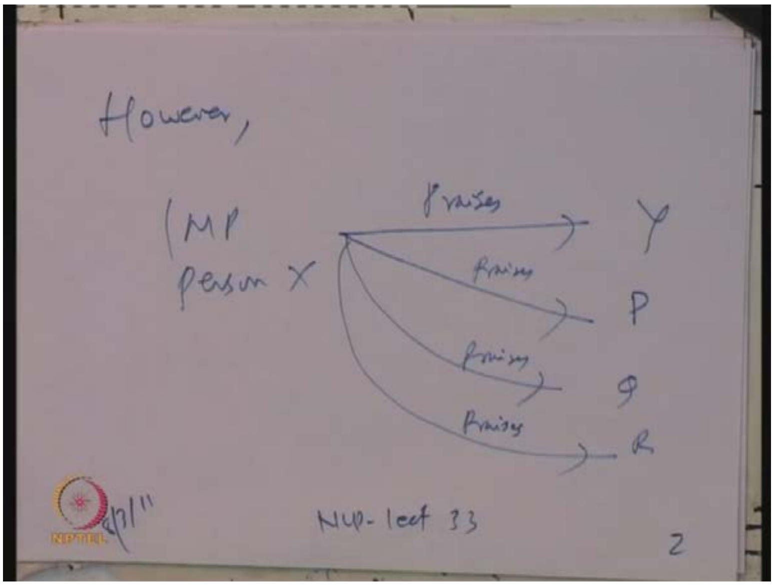
So, this is the generalized formula, and here is the example of its application to its specific situation. Now, what is the basic idea here, let us get the mentuition, first of all we are saying that, the Page Rank of A of A page linking to another pages transfer, and this is transferred wholly, if the page from Page Rank is coming, links to only the receiving page. However, if the Page Rank of confirming note links, to other notes also, then we have to divide this code, into n parts, where n is the number of links coming out from the Page Rank confirming page. What is the idea, the idea is that, suppose an important person, praises another person, so an important person, let us say X praises another person y. So, we would like to, write it down may be, and we will write it will be clear.

(Refer Slide Time: 35:47)



So, person X praises person Y, and this person is not only person, it is an important person. So, since X has praised Y, Y is important also goes up. Since, X has praised Y, Y is important also goes up, this is normal Page Rank transfer.

(Refer Slide Time: 36:50)

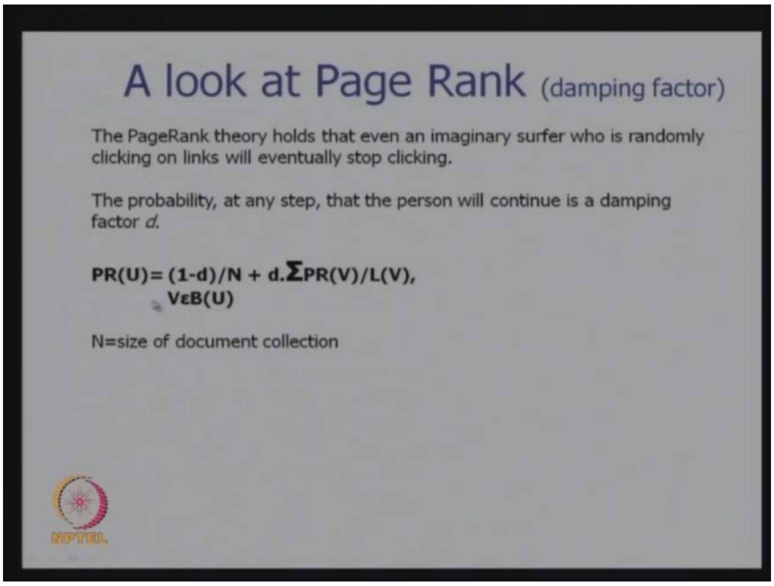


However, if important person X praises Y but he also praises P, praises Q, praises R praised R. Then is not is true, that because this person, is praising, so many people. The importance of his praising, goes down a bit, so one would think that, well person X is important yes, if you selectively gives his praised, to individuals, then there is some weighted in the praise but if the person is owned, to has a habit a praising anybody, then it is doubtful, how much, importance we would be giving, to this praised. How seriously, we should consider this page praise, becomes readability issue, so that why, if you look at the writing here again.

The person's praise goes to Y P Q R, and it will be legitimate of us to say that, the importance, which person X transfers to these people, can be look upon as being divided amongst all of them. If it was only Y, and not PQR then the whole importance of X is attributed to Y but when there were more people, and praises is given to all of them, then we are legitimating thinking, that the importance of X gets starts to all of them equally.

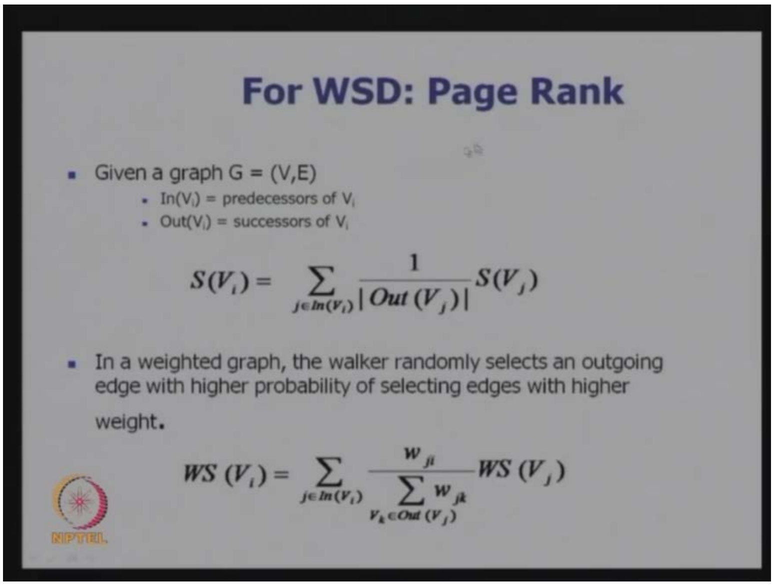
So, then we can say that, transferred importance is equal to importance of X divided by 4, because there are 4 recipients, as you can see Y P Q R. So, this is the main idea, behind the Page Rank algorithm, the Page Rank is transferred from a person to another person. In this case, a note to another note, and if there are many outgoing links, from a note, we divide the Page Rank, and send into the receiving notes.

(Refer Slide Time: 39:59)



So, coming to the slides, now the Page Rank theory holds, that even an imaginary suffer, who is randomly clicking on links, with eventually step clicking, that it true the probability at a, by stop that, the person will continue is a damping factor d. So, that is a bit of adjustment, to the Page Rank algorithm for capturing a phenomenon or person clicking on web pages. So, Page Rank of a document you is PR v sigma PR V by L V, which is basic formula, this is multiplied by the dumping factor d, and 1 minus d multiples 1 minus d e by n, is the other factor. So, this of course, should move to here below, so this is the modified formula for Page Rank.

(Refer Slide Time: 41:01)

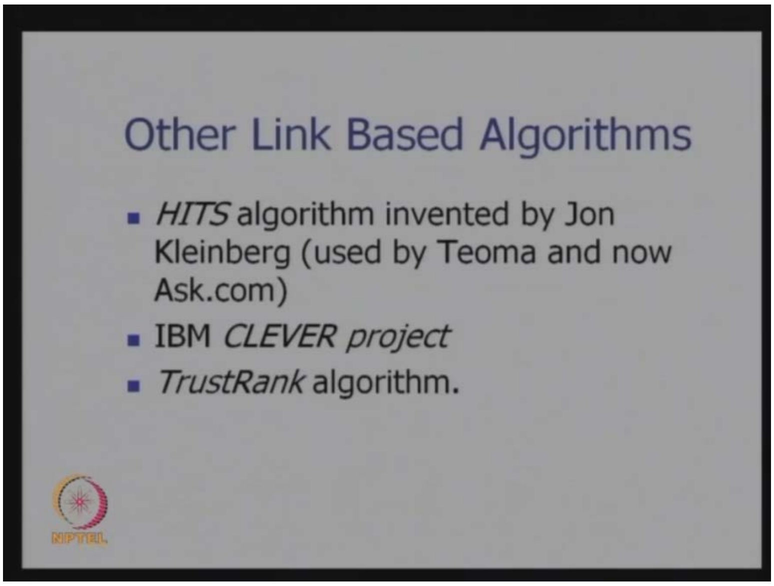


Now, for word sense disambiguation, how does Page Rank apply, so given graph G, V comma E in V i is equal to predecessors of V i, out V i is the successors of V i. So, in a weighted graph, which is what our interest is in, our graph for the senses is a weighted graph. The walker randomly selects an outgoing edge, with higher probability of selecting edges, with higher weight. So, the weighted score of V i is equal to W j i, which is the weight between the note V i and the note V j, divided by W j k, which is for all the outgoing notes or all the out words pages to various notes, and this is summation is over.

All incoming notes, all knows whose arches incoming on V i or incoming on V i, so the picture is really, that of what we was shown before. So, we have this, notes which are the senses, sense notes on the words for this particular .Let us concentrate, on this note, this note will connected to the notes in the next stage, so its outgoing edges from here, to the next words notes sense notes. The incoming links are these, s 3 s 2 s 1 to this particular s 2, so we have incoming links, this and we have outgoing links, form s to and we make use, and we have the weighted edges on this arches. So, now initially, when the Page Rank begins, we have a some Page Rank values randomly, assigned to various notes, and after that by virtue of the Page Rank algorithm.

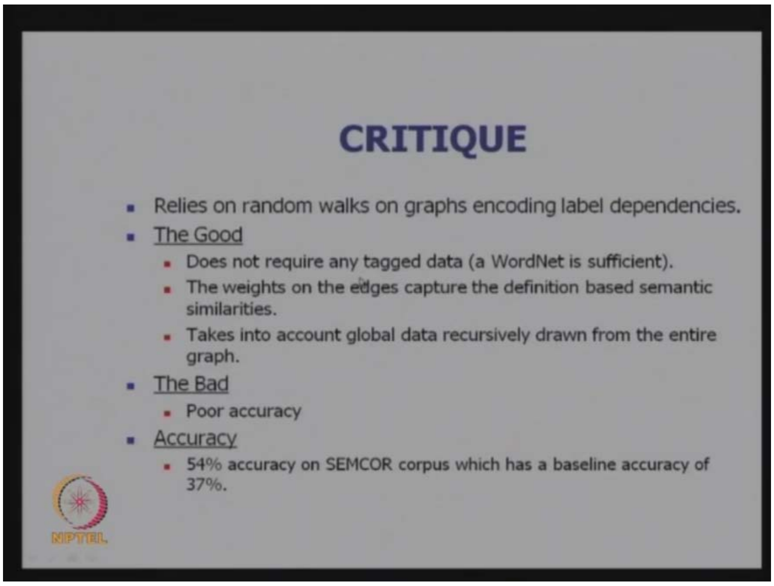
The Page Rank being transferred here for example, e is the as per the weighted here, weighted on this, and divided by the number of links weighted of the links, coming from a s 3. So, this is the Page Rank transfer to s 2, and from here again Page Rank gets transferred to all these senses. So, this the Page Rank here also get modified, so this algorithm goes on running, and after sometime there is a conversions, and when the conversions happens, we get the winner senses.

(Refer Slide Time: 44:23)



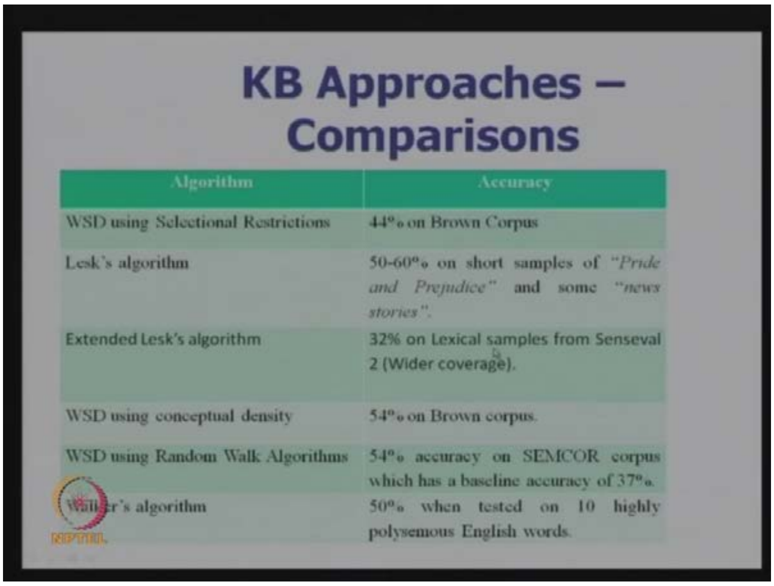
Then other based algorithms, which are used for ranking of web pages, hits algorithm was invented by Jon Kleinberg used by Teoma, and now Ask.com, IBM CLEVER project. And, then there is also, mething called the Trust Rank algorithm, these algorithms have been not been tried for word sense disambiguation.

(Refer Slide Time: 44:47)



So, the critique, so the Page Rank algorithm, Page Rank based algorithm is that, it relies on random walks ‘operation graphs encoding label dependencies, The Good part is that, it does not requires any tagged data, a Wordnet is sufficient, the weights on the edges, the definition based semantic similarities, takes into account global data, recursively drawn from the entire graph. And the bad is that, in spite of its attractive properties, and the attractive nature of the algorithm, the algorithm has a poor accuracy. So, this is something which is worth investigating, the results were known of course, but as a student it will be useful to you, to get inside in the algorithm, and understand why the algorithm, has poor accuracy, so this alike the conceptual density idea it, how about 54 percent accuracy on SEMCOR corpus which has a baseline accuracy of 37 percent.

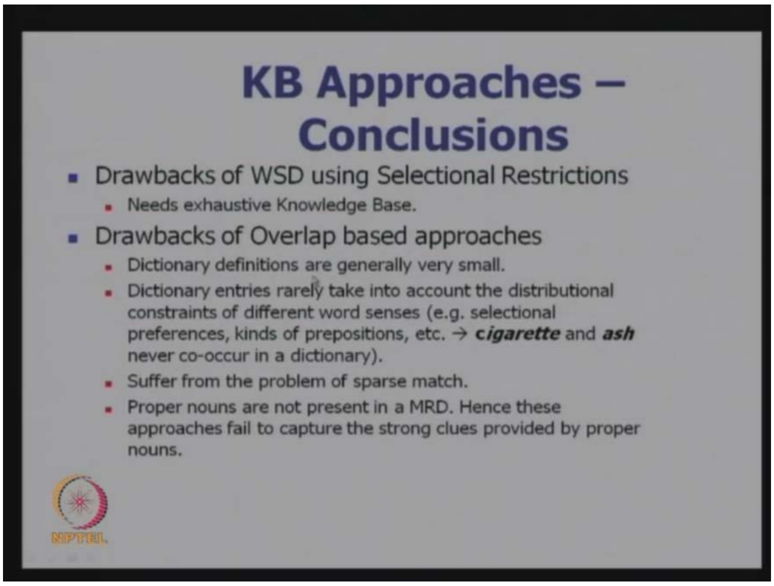
(Refer Slide Time: 45:57)



So, this would finish the Knowledge Based Approaches, and we can make some summarizing remarks here, on the Knowledge Based Approaches to word sense disambiguation, word sense disambiguation in which selection restriction, where the arguments and the properties are considered. It has an accuracy of 44 percent on brain brown corpus, Lesk's algorithm, which is overlap based 50-50 to 60 percent on short samples, of pride and prejudice and some news stories. Extended Lesk's algorithm; it has an accuracy of 32 percent on lexical samoles, from senseval 2, well the accuracy figure becomes slower, because of higher troop tropic drift, word sense disambiguation, using conceptual density.

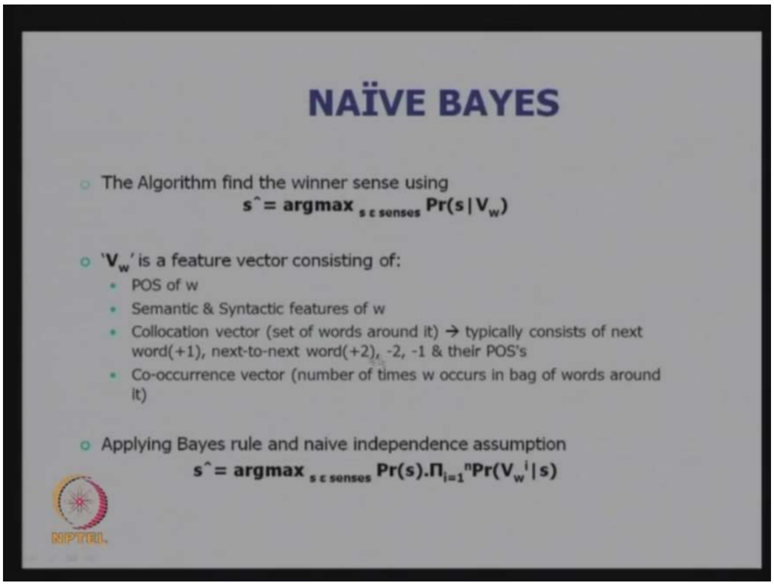
It has an accuracy of 54 percent on the brown corpus, word sense disambiguation using random walk algorithm, 54 percent of accuracy on semcor corpus, which has a baseline accuracy of 37 percent. Walker's algorithm produces a 50 percent accuracy, when tested on 10 highly polysemous English words. Now, you see this accuracy figures, are reported, but they cannot really be compared, the reason is that we are not using the same data, for evaluation. If the same data was, use for evaluation then they were comparable but these are only these are indicative figures, which tell us, what kind of accuracy, what can whom for when one uses these algorithms.

(Refer Slide Time: 47:36)



Proceeding further, we look at the drawbacks of WSD, which uses the Knowledge Based Approaches, here the drawbacks of WSD using sectional restriction are, they needs exhaustive knowledge bases, all the properties of above nouns have to be stored. Drawbacks of overlap based approaches is that the dictionary definitions, are generally very small dictionary, entries rarely take into account, the distributional constraints of different word senses for example, selectional preferences, kinds of prepositions etcetera. Cigarette and ash never co occur in a dictionary, they suffer from the problem of spares match, proper nouns are not present in a machine readable dictionary, hence these approaches fail, to capture the strong clues provided by proper nouns. Now, we move on to supervised approaches, and we will just start it, to be elaborate in the next lecture. In supervised approaches, the whole idea is to have, sense mark corpus, and we learn from the sense mark corpus, so if you look at this slide.

(Refer Slide Time: 48:52)



Here, the first important approach to supervised approaches is the Naïve Bayes algorithm, this conducted through an argmax, based computation, I will just put down the formula, and make an some initial remarks on this. The winner sense, is obtain by computing the probability of the sense given, the target words features V w is the target words feature vector. And we do it for all possible senses, and the sense, which has the best argmax score, becomes the winner sense. The feature vector, consist of part of speech, semantic and syntactic features, collocation co occurrence vector etcetera. So, we will discuss this algorithm, and see how sense mark corpus can be used for sense disambiguation.