Natural Language Processing

Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

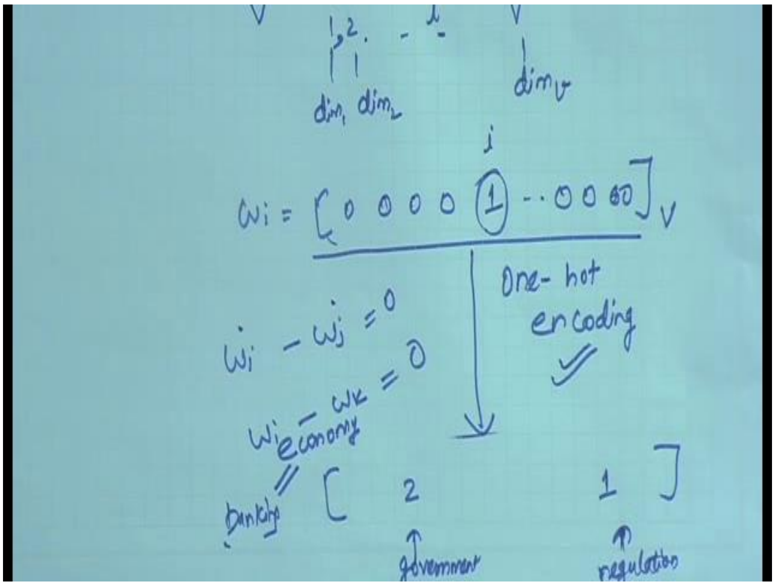
Lecture – 32
Distributional Models of Semantics

Welcome back for the second lecture of this week. So, in the last lecture, we started our discussions on distribution semantics and we took a very simple example in the very naive method we constructed that distribution matrix and try to compute the similarity across words. Now you will be see this concept much more formally that how do we construct such models and what are the different applications that that we can use them all, we will see some examples.

Firstly, let me just go back a step and see why do we need to talk about the distributional semantics? So, if I talk about a vector space model without distributional semantics, what would happen? So, there words would we treated as various atomic symbols. So, what do I mean by that?

Suppose think of your semantic space as corresponding to various different words that you see in your vocabulary.

(Refer Slide Time: 01:19)



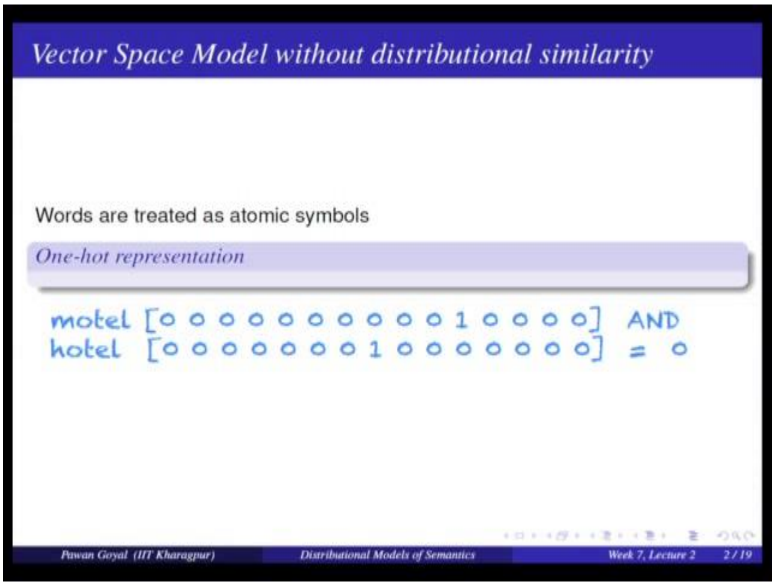
If your vocabulary is of size V, so, contains word with index 1, 2 up to V and I want to give a representation to these words in the vocabulary some representation. So, what is the representation that I can give without looking at any distributions? So, I will say that each one is a different dimension; dimension 1, dimension 2, up to dimension V and how do I denote a word in these dimensions?

Suppose I have the i th word. So, I will say, so let me denote it the i th word, this will be a vector in V dimensions where only the i th element is 1, everything else is 0. So, like that I can give a different representation to each of the V words where only 1 element corresponding to the index will be 1, everything else will be 0 and this is also called the one-hot encoding, only one of the entry is 1, everything else is 0.

Now, so there is also my semantic space where all the V dimensions correspond to different V different words. Now what is the difference here? I am not capturing distribution, I am just saying this word is the i th dimension nothing else. Now can I do semantics with that or can I capture the meaning of the words with that? Each word has a different vector, now suppose I want to find out if w i and w j are more similar than w i and w k. So, what is the similarity between w i w j and w i w k?

What will happen if I use this encoding? So, I will find out because here the i th element will be 1, everything else will be 0, here j th element will be 1, everything else will be 0. If I try to take a dot product, I will get a 0 in both the cases, even if you find that these 2 words are looking similar than these 2 words and that is 1 problem with using on the one hot encoding.

(Refer Slide Time: 03:39)



Vector Space Model without distributional similarity

Words are treated as atomic symbols

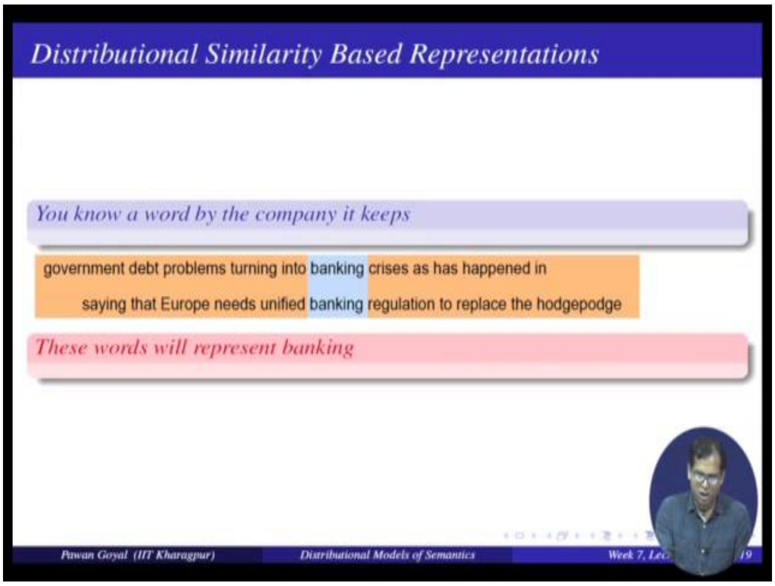
One-hot representation

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2 2 / 19

Suppose I take 2 words here like motel and hotel and suppose motel occur at this index, it will has 1 everything as 0, hotel occurs this syntax, this is 1 everything as 0. So, if I take to take try to take the dot product of these 2, I will get a 0 and if I take hotel and book that also will give me a dot product of 0. So, this will not capture that hotels and motels are much more similar than hotels and book.

(Refer Slide Time: 04:10)



Distributional Similarity Based Representations

You know a word by the company it keeps

government debt problems turning into banking crises as has happened in
saying that Europe needs unified banking regulation to replace the hodgepodge

These words will represent banking

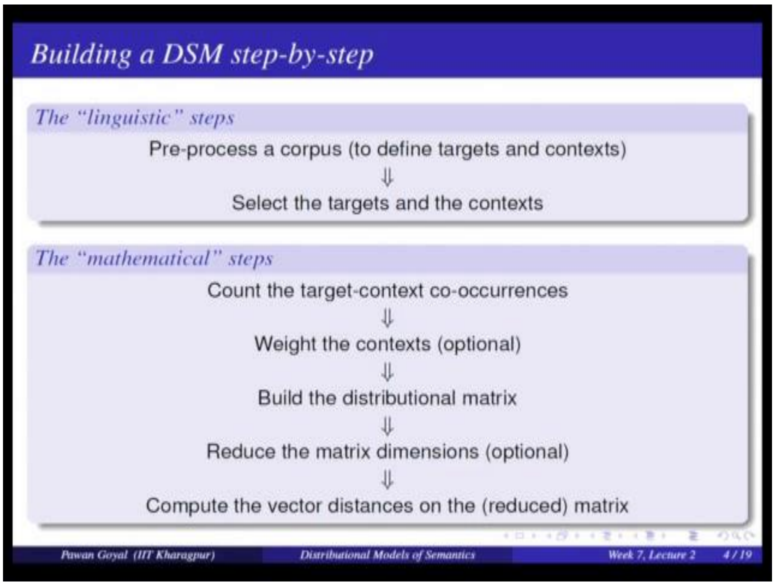
Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lec 19

Now what do we do in distribution semantics? We use this idea that you know a word by the company it keeps. So, instead of putting only 1 and everything else 0, try to model the distributions. So, suppose I take a word like banking, I will not denote it by a simply one at its syntax, but I will see what are the other words that come its context.

If we talk it is occurring in the corpus with government, dept, problems, turning crises, Europe, regulations, replace, etcetera, then these words will be use to represent banking. So, now, what will happen? So, instead of my one hot encoding, I will now go to a distribution representation where I will say for the word like banking, how many times suppose this is my dimensions corresponding to government.

How many times it occurs with government? Say 2 times, similarly here it is my dimension corresponding to say regulation, how many times it is occurring with regulation? So, this is I am doing for bank, now I can do similar thing for something else like economy and so on, what they will find? Probably the word banking economies have quite similar distributions than the word like banking and something like play, yes. So, instead of having one-hot encoding I am having a distributed representation and that is what is being done by these distribution semantic models.

(Refer Slide Time: 05:58)



Building a DSM step-by-step

The "linguistic" steps

Pre-process a corpus (to define targets and contexts)
⇓
Select the targets and the contexts

The "mathematical" steps

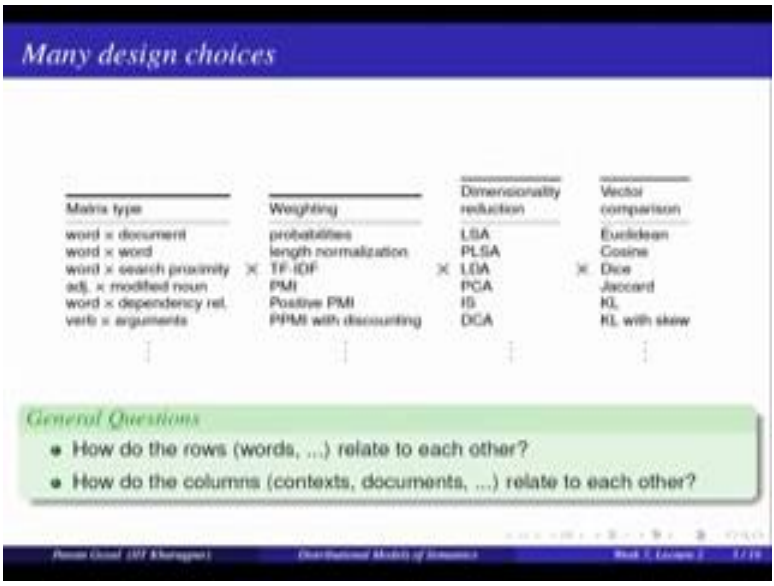
Count the target-context co-occurrences
⇓
Weight the contexts (optional)
⇓
Build the distributional matrix
⇓
Reduce the matrix dimensions (optional)
⇓
Compute the vector distances on the (reduced) matrix

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2 4 / 19

Now, how would I build a distribution semantic model step by step? So, you will take a corpus that is the that is something that you have require that contains a lot of data the various usage of different words in the sentences now you will do some pre processes to define, what are your target words and what are your context words? Context can be context word or something else, we will see some examples. Now once you select what are your targets and context from your data and what is the next task.

Next you will count, how many times the target co occurs with various context and you will give some weights to these contexts it can be the simply number of times they co occur or some function applied over this number and we will see again see some examples for that. Now once you have given the weights, you build the matrix; distribution matrix now some optional step is if you want to reduce the dimensions of this matrix because the dimensions in general can be very very high, if you are taking talking about a large corpus, you might have millions of words as your context. So, do you want to further reduce the dimensions, then once you have the representation in that reduce dimension you can also try to capture the similarity across words. So, you compute the vector distances on the reduced matrix and this is the overall pipeline for constructing a distribution semantic model.

(Refer Slide Time: 07:28)



Many design choices

Matrix type		Weighting		Dimensionality reduction		Vector comparison
word × document		probabilities		LSA		Euclidean
word × word		length normalization		PLSA		Cosine
word × search proximity	×	TF-IDF	×	LDA	×	Dice
adj. × modified noun		PMI		PCA		Jaccard
word × dependency rel.		Positive PMI		IS		KL
verb × arguments		PPMI with discounting		DCA		KL with skew
⋮		⋮		⋮		⋮

General Questions

- How do the rows (words, ...) relate to each other?
- How do the columns (contexts, documents, ...) relate to each other?

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2

.

Now, here there are many design choices that you might have to make, for example, we said that we will define what are my target words, what are my context words now it can be I there can be many many variations. So, symbol is could be my targets are words and contexts are documents and I am saying how many names this word occurs in this documents and further I can give various weights.

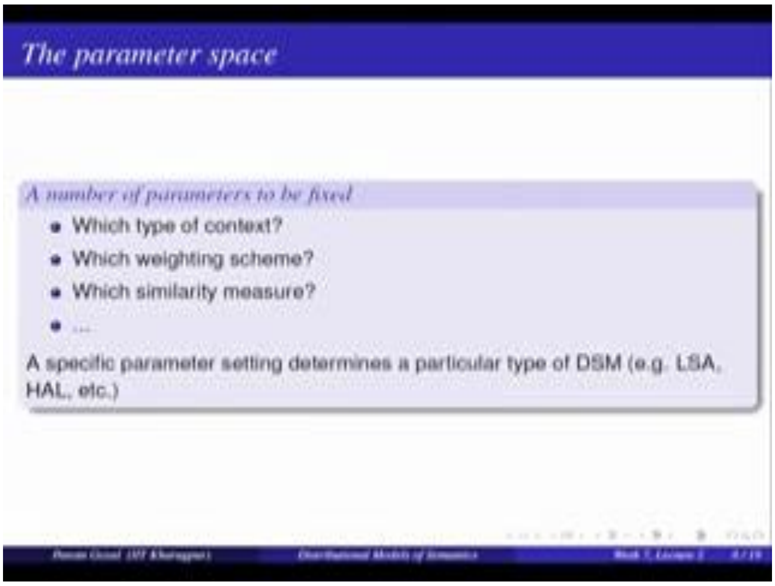
This is 1 matrix type, another could be word cross word how many times this word occurs with another word then again keep on going. So, how many times this word occurs with that word in certain such proximity or this adjective occurs with this modified noun and word with dependency relation noun and verb with its argument and so on. So, I can have various different sort of matrix types initially we will only focus on the first and second that is my word is the target and document is the context or word is a target and word is the context.

Now, once I have this matrix, what are the other design choices? Then how do I weight the elements? Should I use the probabilities to weight them? Should I use length normalization? TF-IDF, PMI, positive PMI or positive PMI with discounting, what is the method that I should use for weighting the entries in my matrix? Then if I am doing dimensionality reduction, there are the many methods LSA, PLSA, LDA is also one sort of dimension reduction method PCA so on, then once even I have done that how do I compare 2 different vectors?

Once I have built all the vectors in this manner, should I use the Euclidean distance, cosine similarity or dice coefficient, Jaccard coefficient, etcetera. So, they are lot of design choices that you might have to make and each individual on might make a different sort of distribution semantics model, but the underlining idea is the same that you want to capture a distributions from a large corpus.

Now and then there might be different questions that you might be going to answer that once you have built the distribution matrix how do different rows in the matrix, let to each other and how do different columns in the matrix let to each other.

(Refer Slide Time: 09:48)



The parameter space

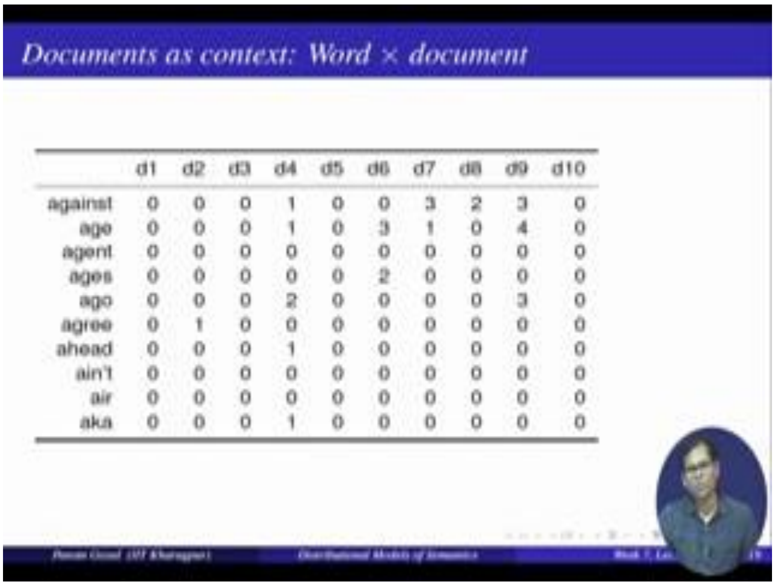
A number of parameters to be fixed

- Which type of context?
- Which weighting scheme?
- Which similarity measure?
- ...

A specific parameter setting determines a particular type of DSM (e.g. LSA, HAL, etc.)

What we have seen a number of parameters that we need to fix like what type of context should I use? What weighting schemes are used? Yes and what similarity measures should I use and different models might be different settings of these parameters n now.

(Refer Slide Time: 10:07)

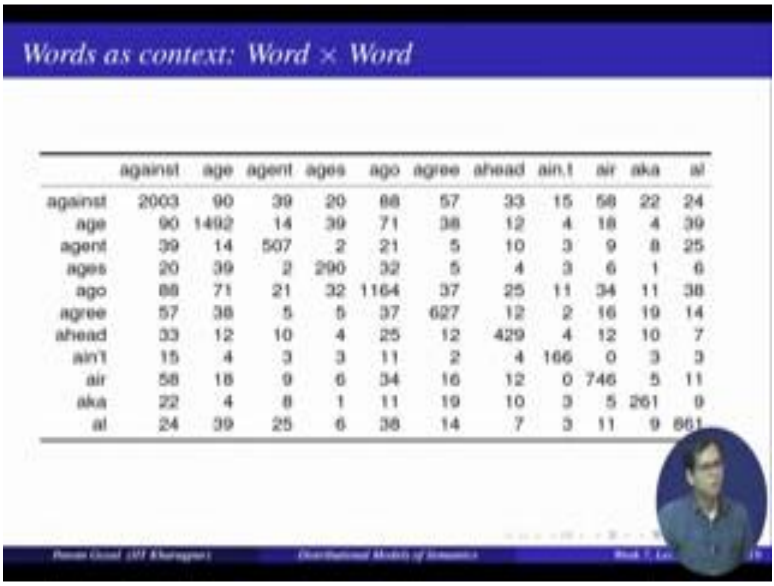


Documents as context: Word × document

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
against	0	0	0	1	0	0	3	2	3	0
age	0	0	0	1	0	3	1	0	4	0
agent	0	0	0	0	0	0	0	0	0	0
ages	0	0	0	0	0	2	0	0	0	0
ago	0	0	0	2	0	0	0	0	3	0
agree	0	1	0	0	0	0	0	0	0	0
ahead	0	0	0	1	0	0	0	0	0	0
ain't	0	0	0	0	0	0	0	0	0	0
air	0	0	0	0	0	0	0	0	0	0
aka	0	0	0	1	0	0	0	0	0	0

Let us start with the simplest case where words are the targets and documents are the context. Now here is a simple example. So, you are seeing some words as a rows and documents as columns and what do the entries denote that against occurs in the document d 4 1 times in d 7 3 times, d 8 2 times d 9 3 times, as per the document it occurs, it does not occurs and that is how you are represent giving a representation to the word against these doing without any weighting just simply the row counts and this you can be very easily find for any different any word. So, in information table also this is done for each word how many times it occurs in various documents.

(Refer Slide Time: 10:52)



Words as context: Word × Word

	against	age	agent	ages	ago	agree	ahead	ain,t	air	aka	al
against	2003	90	39	20	88	57	33	15	58	22	24
age	90	1492	14	39	71	38	12	4	18	4	39
agent	39	14	507	2	21	5	10	3	9	8	25
ages	20	39	2	290	32	5	4	3	6	1	6
ago	88	71	21	32	1164	37	25	11	34	11	38
agree	57	38	5	5	37	627	12	2	16	19	14
ahead	33	12	10	4	25	12	429	4	12	10	7
ain't	15	4	3	3	11	2	4	166	0	3	3
air	58	18	9	6	34	16	12	0	746	5	11
aka	22	4	8	1	11	19	10	3	5	261	9
al	24	39	25	6	38	14	7	3	11	9	861

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Le

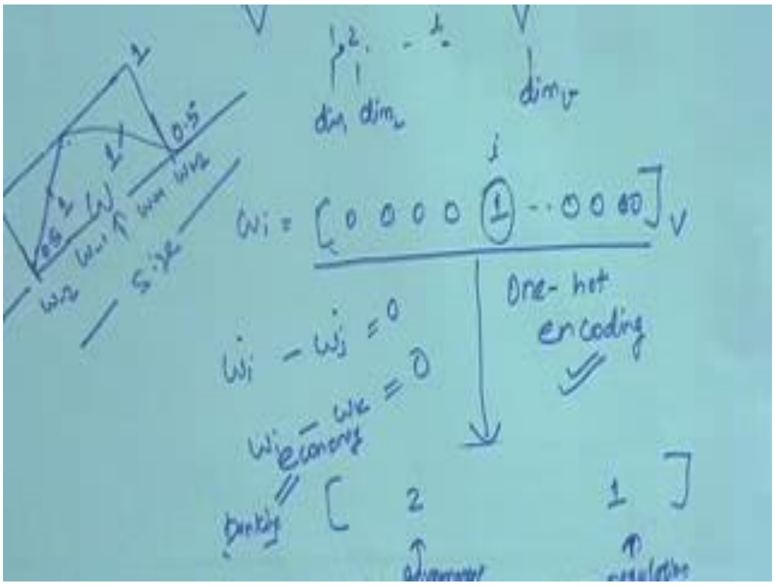
.

On the other hand suppose your contexts are the word then you are trying to show how many times a word occurs with another word. So, in this particular case, so the word against occurs with age 90 times, age agent 39 times and so on and this diagnoses tells how many times the word agent occurs. So, it is being computed by as if co occurring with itself. So here so, again you will define a context window, it can be a sentence or something and you will find out, how many times the word air occurs with against and age occurs with age and so on, this you can do for all the words or a specific set of word that you have already chosen for your analysis.

This is simply the count, now if I am using my words as contexts; document contexts are easy because I know, what is the context window size? This is the whole document. So, I will say whether these words occur in the document or not or how many times it occurs? Now suppose I am using words at constant now, now the question comes in what should be the size of my context window should I choose only the co occurrence within 2 words occurring or within a paragraph, within the whole document and this is the design choice that you can make, depending on are you trying to measure similarity on based on some very close co occurrences or you are alloying even a some very very distant co occurrences in the same document. So, what is the parameter here? What is a size of my window? So, how far am I looking it and what is the shape of my window? Is it rectangular, triangular or something else? Let me just briefly let what do I mean by the shape of the window.

So, that is suppose I am having a word w.

(Refer Slide Time: 12:50)

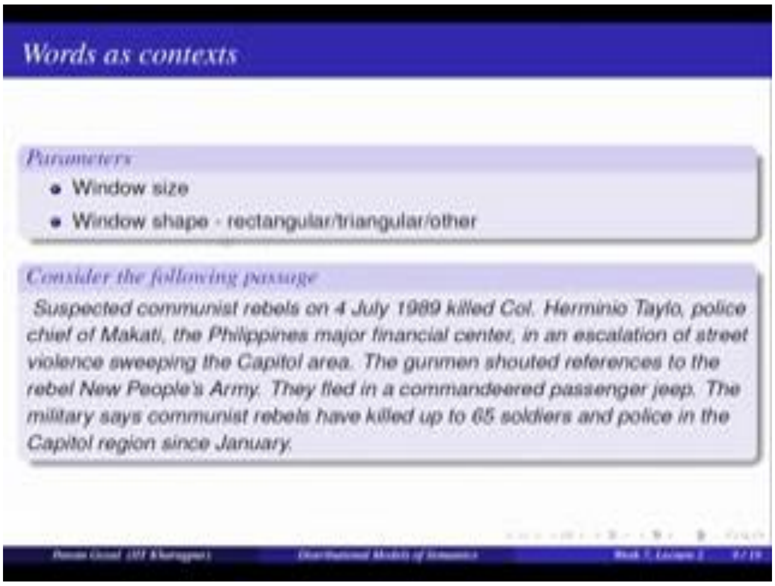


And it has some context. So, I am trying to see how many times this word occurs with these words around the context. So, suppose this is the previous word, previous to previous word, next word, next to next word, so my size is am I looking at only 2 words around it or more than that 3, 4 and so on, this is my size. So, this is the size and what do I mean by shape? So, we said it can be triangular. So, triangular will be something like this is rectangular, what is the difference in triangular?

What I will do as I keep on moving away from my target word, I reduce the count, I will say this count, I will treat as 1, this I will treat as 0.5. So, this is the strength is decrease as you go away from the target word similarly this can be one this can be point 5 and you can as such were give it any shape like exponential or whatever, in that that way you can capture even much further co occurrences in that (Refer time: 14:13) you know what you will what you will do you will say I will count each co occurrence size same way. So, each has a weight of 1 or whatever you give.

This is the idea of window size and window shape.

(Refer Slide Time: 14:27)



Words as contexts

Parameters

- Window size
- Window shape - rectangular/triangular/other

Consider the following passage

Suspected communist rebels on 4 July 1989 killed Col. Herminio Taylo, police chief of Makati, the Philippines major financial center, in an escalation of street violence sweeping the Capitol area. The gunmen shouted references to the rebel New People's Army. They fled in a commandeered passenger jeep. The military says communist rebels have killed up to 65 soldiers and police in the Capitol region since January.

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2

Now let us take 1 example from some news article. So, this is my passage the suspected communist rebels on 4th July 1989 killed and so on, this is my passage and I want to capture the co occurrence over there are. So, what I will do? I will first define, what is my window size? Suppose I have a window of size 5, I want to use window of size 5, where I am taking 2 words either side of my target word.

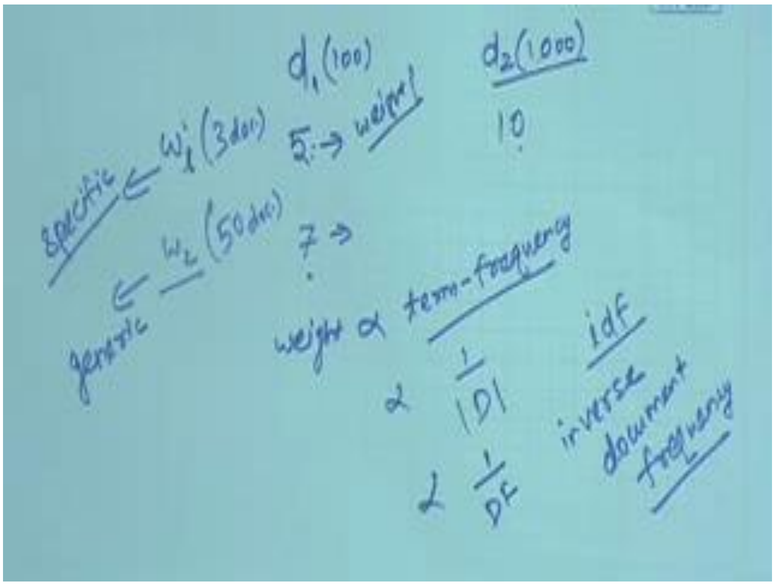
What will the window look like? Take the word like rebels here I will find out its co occurrence only with 2 words on the left and 2 words on the right. So, here suspected and communist come here on and 4 come here, similarly when rebels occur here says and communist come here, have and killed come here. So, these are my context words in the window.

I will only define rebels coming with these words, everything else I will not, I will count as 0. So, this is my unfiltered window, I am not filtering any (Refer Time: 15:43) words, I can also use a filtered window where I can filter the (Refer Time: 15:48) words. So, something like that. So, I will take 2 words either side of the target word, but now these are filtered. So, that they do not contain the stop words.

Here on the left I have the 2 words suspect and communist, but on the right I remove the words on 4 and a number like 1989 and I am taking the words July and killed similarly here I have says and communist in my window, but not have is like a stop word. So, I have removed that and then killed is there up to 65 is not there and soldiers. So, you can take either the filtered window or an unfiltered window.

Now, how do I weight the context? So, again let us take the 2 cases when my documents are context was is when my words are context. So, when my documents are context how should I weight the occurrence of a word in my document? So, what it should depend on? So, in how many times a word is occurring in the document? What it should depend on or what the weight should depend on?

(Refer Slide Time: 16:59)



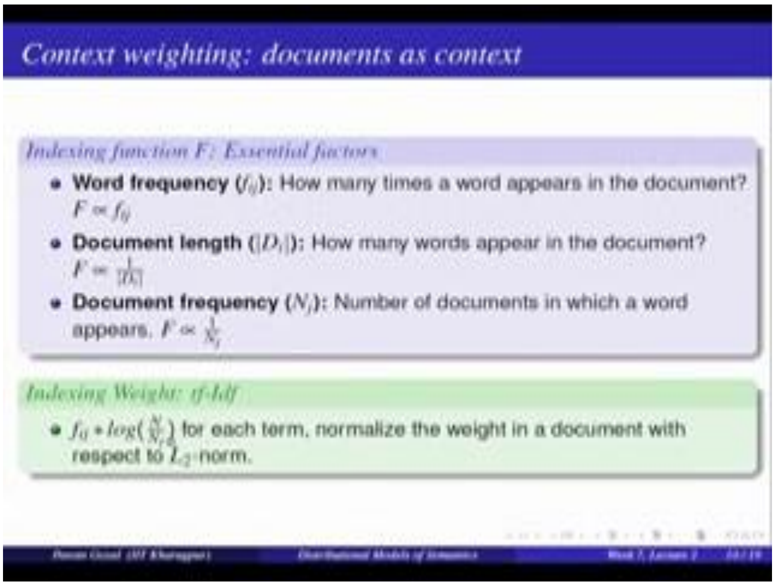
I am denoting a word in documents and I am going to give a weight suppose it occurs 5 times in this documents and another word occurs seven times and now I am convert that into some weight. So, what should this weight depend on? So, what can be the different parameters on which it can depend? Firstly, if a word occurs more number of times, it should get a high weight. So, this weight should the proportional to number of times, it occurs and we call it the term frequency how many times a word occurs in a document.

What else if the document is very very long. So, you might have take that window count. So, suppose I have a document d one of size 100 and I have a document d 2 of size thousand and I know that word one occurs in d one with 5 times and it occurs in d 2 ten times I cannot simply say that in document 2 word, 1 has a high weight because document 2 itself is very very long. So, the weight should be inversely proportion to the length of the document as the length of the document increases the weight should also reduce and what can be the other measure another measure is very very important also you can you can even know that if you have heard about some information debate topics.

This is called the inverse document frequency. So, that is IDF inverse document frequency and what is the idea. So, idea is if a word occurs in many many documents through my corpus was such if a words occurs in only selective documents whom should I give a more weight. So, that is suppose my word w one occurs in this document 5 times and w 2 occurs seven times that is, but suppose I also find that w 1 occurs only in 3 documents overall and w 2 occurs in 50 documents.

What can I say about the nature of these 2 words? So, one thing I can say is that w 2 is probably a very generic term and w one is a very very specific term this becomes a generic term and this is a specific term. So, idea is that specific term might have more information about the document than a generic term because generic term might be occurring in many many different documents. So, I want to give a high weight to w one because it occurs in very few documents and it has occurred here so; that means, I will give a weight proportional to one divide by number of documents that it occurs in. So, I can call it a document frequency if the document frequency is high I give it a small weight.

(Refer Slide Time: 20:38)



Context weighting: documents as context

Indexing function F: Essential factors

- **Word frequency** (f_{ij}): How many times a word appears in the document? $F \propto f_{ij}$
- **Document length** ($|D_i|$): How many words appear in the document? $F \propto \frac{1}{|D_i|}$
- **Document frequency** (N_j): Number of documents in which a word appears. $F \propto \frac{1}{N_j}$

Indexing Weight: tf-Idf

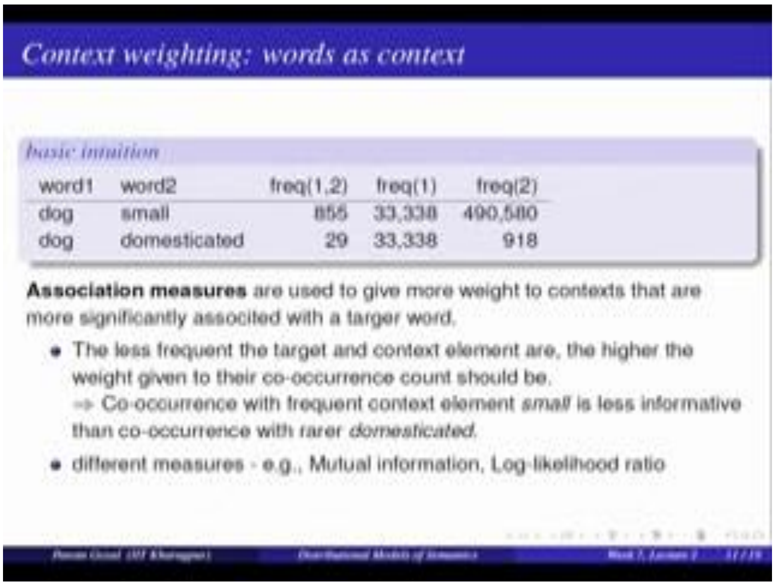
- $f_{ij} * log(\frac{N}{N_j})$ for each term, normalize the weight in a document with respect to L_2-norm.

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2

These are 3 different parameters or factors on which my function bit will depend on how many times a word appears in the document then what is the different number of words that occur in a document and how many different documents a word occurs in and as we saw, its weight should be proportion to the number of times that occur in the documents should be inversely proportion to the other 2 factors and there are various different induction functions that that have been proposed that take into account all these ideas and one very common measure that is used a TF-IDF that is I give a weight of proportional to the frequency F i j times log of n by n j. So, n j is the document frequency. So, if it is higher, I give it a smaller weight and where t is the document coming in to picture document length. So, because once I compute the t f IDF for each individual term, the document I take the l 2 norm. So, if there are lots of terms, the document, the individual weights will be reduced.

This is 1 very commonly used induction function called TF-IDF I see what are the number of times that word occurs in the document and how many different documents that occurs in. So, 1 thing, they have many variations that have been proposed for this function this is the simplest function that has been know for TF-IDF and finally, we have to take the l 2 norm, this is something that we should not forget that I have to finally, normalize all the different values in a single document. So, that some of the squares adds up to 1. So, now, suppose I take words as my context instead of documents then what is the interesting weighting function I can use?

(Refer Slide Time: 22:33)



Context weighting: words as context

basic intuition

word1	word2	freq(1,2)	freq(1)	freq(2)
dog	small	855	33,338	490,580
dog	domesticated	29	33,338	918

Association measures are used to give more weight to contexts that are more significantly associted with a targer word.

- The less frequent the target and context element are, the higher the weight given to their co-occurrence count should be.
  ⇒ Co-occurrence with frequent context element *small* is less informative than co-occurrence with rarer *domesticated*.
- different measures - e.g., Mutual information, Log-likelihood ratio

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2 17 / 19

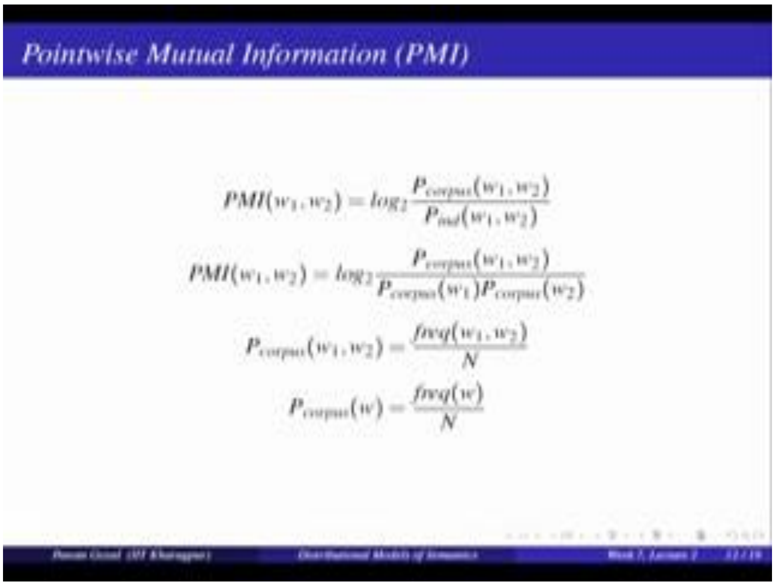
Let us take an example. So, here my target words is word is dog in both the cases context word is small and domesticated 2 different words as my context and what you been shown here. So, target word has the same frequency yes, but the context words in one case is small as if it sorry small has a frequency of 490580 and domesticated has a frequency of 980. So, you can clearly see here see here that the word small is very very common and domesticated is very specific term.

Now, suppose I find out the co occurrence the dog are occurs with small 855 times and dog occurs with domesticated 29 times, now my question is what number I should used denote, how much dog co occurs with the word like small and domesticated. So, if I do not give any weights what will happen here I will say dog occurs a lot with the small 855 times and dog occurs very rarely with domesticated 29 times, but that is not capture the whole picture.

What is the whole picture domesticate is the very very rare word, it occurs only in 900 documents out of which in 30 documents occurs with dog a small occurs in 490000 documents out of which only in 855 documents, it occurs with dog. So, can I use the idea that if a word is rare word, its co occurrence with the target word should have higher weights than if the word is very very common word? So, that is can I use the frequency of the individual words also when I give the weights to their co occurrence and that is what we do why using various association measures that I used to give various weights to the context and the idea is that the less frequent target and context elements are the higher the weight you give to their co occurrence count.

In this case, what will happen? So, because my context element is less frequent here in domesticated their occurrence with dog should be given a high weight. So, co occurrence with the frequent context element small will be less informative here than the co occurrence with the rarer word domesticated and there are various measures that captures that like mutual information is very very popular measure and also log likelihood ratio etcetera they try to capture this. So, how common and rare or rare the words that among which I am finding co occurrence are and how many times they co occurred together. So, they try to use both of these into a single function.

(Refer Slide Time: 25:12)



Pointwise Mutual Information (PMI)

$$PMI(w_1, w_2) = log_2 \frac{P_{corpus}(w_1, w_2)}{P_{ind}(w_1, w_2)}$$

$$PMI(w_1, w_2) = log_2 \frac{P_{corpus}(w_1, w_2)}{P_{corpus}(w_1)P_{corpus}(w_2)}$$

$$P_{corpus}(w_1, w_2) = \frac{freq(w_1, w_2)}{N}$$

$$P_{corpus}(w) = \frac{freq(w)}{N}$$

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2 12 / 19

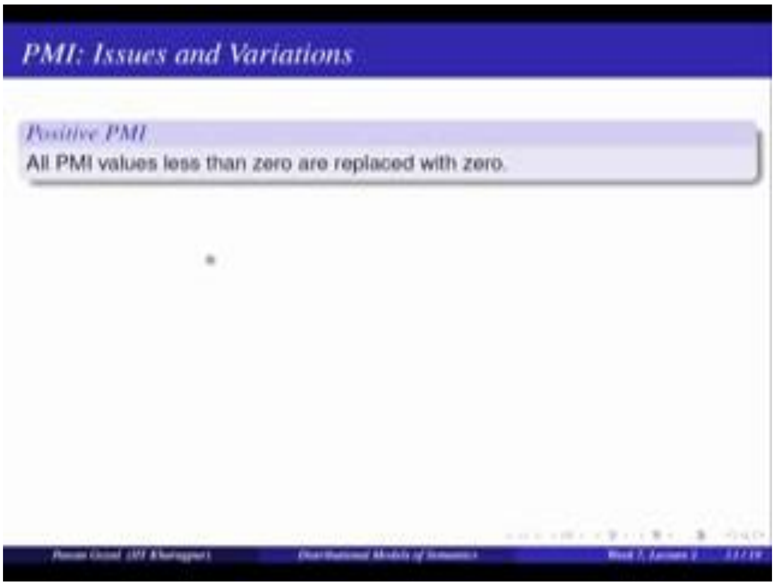
What is my, what is the function for point wise mutual information for 2 different words; w 1 w 2? I will find out in mutual information.

What is the probability that they co occurred together in the corpus? I will divide it by what is the probability that they would have co occurred together had they been they been occurring independently. So, this mutual information tells me, how much information do I gain by seeing? How many times they are co occurring that I would not have obtained by their random co occurrences. So, that is the formula. So, what is their probability of they co occurrence in the corpus divided by what is the probability of a co occurrence in the corpus if they were independent now how do we capture the probability of their occurring of the occurring independently in the corpus.

That would be I will say the probability with which w one occurs and among those times w 2 could have occurred again with this probability. So, it will be the probability of occurrence of w one times probability of occurrence of w 2. So, so this is a function that I can use probability with which w 1 w 2 co occur divide by probability with which w 1 occurs times probability with which w 2 occurs and I will take a log of for that and how do I capture probability of occurrence in a in the corpus for 2 words number of times they co occurred together is a simple co occurrence count divided by n.

N here would denote the different number of context that you can see or in or it can be also the number of different tokens that you have in your data similarly probability corpus of the word w is very simple this is the unigram model number of times this word occurs divided by a all total different number of token that you have seen.

(Refer Slide Time: 27:19)



PMI: Issues and Variations

Positive PMI

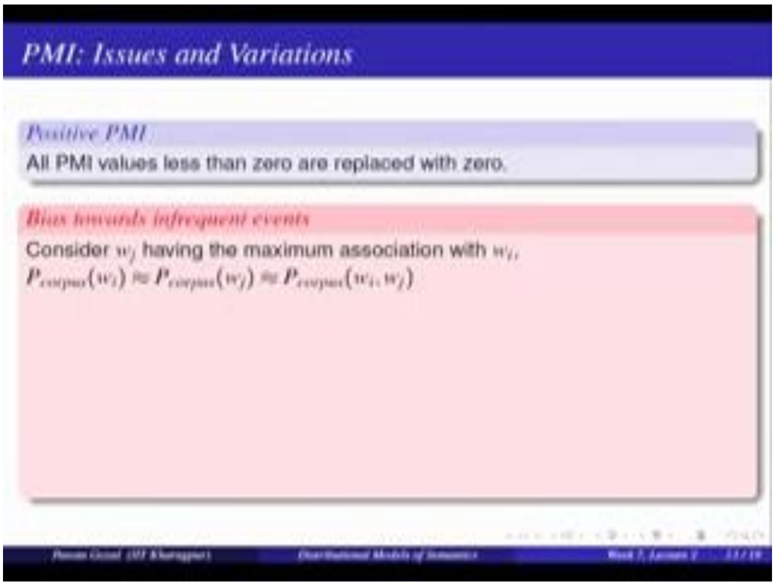
All PMI values less than zero are replaced with zero.

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2 11/19

.

In PMI, what would happen in sudden cases? The value might also go to negative values. So, what we what is some variations there? So, you can use up only the positive values and this is called positive PMI, PPMI. So, where only those values that are greater than 0 are taking into consideration everything else is converted to 0.

(Refer Slide Time: 27:46)



PMI: Issues and Variations

Positive PMI

All PMI values less than zero are replaced with zero.

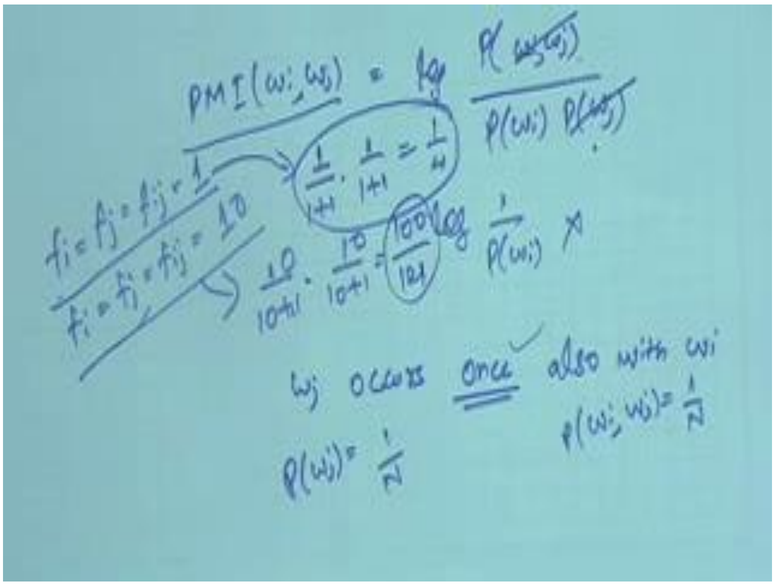
Bias towards infrequent events

Consider w_j having the maximum association with w_i,
$P_{corpus}(w_i) \approx P_{corpus}(w_j) \approx P_{corpus}(w_i, w_j)$

Pawan Goyal (IIT Kharagpur) Distributional Models of Semantics Week 7, Lecture 2 11/19

Now, there is one problem in this mutual information approach that there is a biased word some in frequent evens. So, remember why we came this idea of mutual information we wanted to give high weight to the events that are in frequent if 2 words are rare, we do wanted to give their association a high weights, but what happens if 2 evens 2 words are very very rare they get they might get some high weight. So, some one particular cases suppose think about the scenario with w i and w j have the same occurrence in the corpus as w i w j w j together now what to do the probability or the mutual information in this case.

(Refer Slide Time: 28:32)

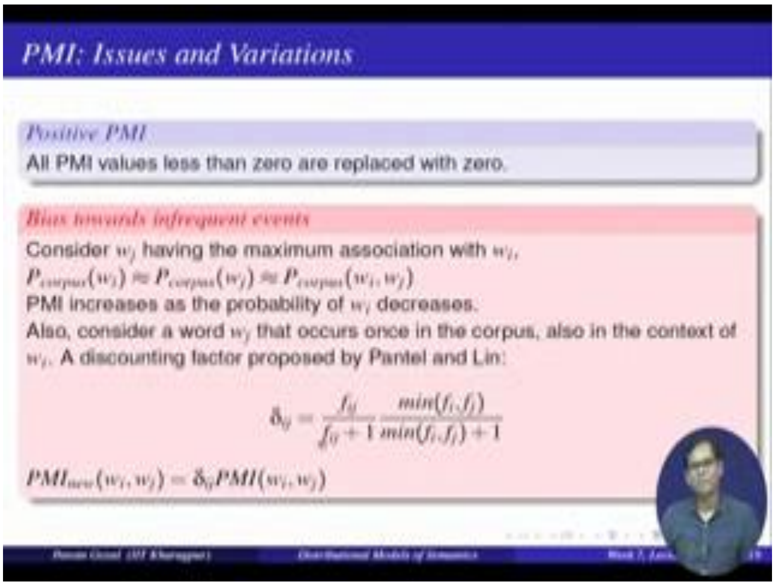


So, PMI between w i w j is log of probability w i w j divided by probability w i times probability w j and suppose all 3 are equal. So, what would happen? This would cancel. So, this will be log 1 by p w i. So, what it is saying if these 3 are equal PMI will be high, if PMI is low and immediately that that gives a bias in the favor of events that are smaller though or words that occur very very rarely. So, you can think of 2 scenarios where all this 3 probabilities are 1 by 100 or versus all these 3 probabilities are 1 by 10000. So, either we I would like the association to be the same in both the cases, but what would happen in this formula the association become high when the probability is low and this is not desired. So, this will create a problem when their individual probabilities are very very low, this immediately will become PMI will become very very high.

And there are another case where you can see that if the word w j occurs only once in the corpus and it also occurs that time in with w i. So, what would happen? So, P w I w j and P w j will become similar here also because what I am I saying w j occurs once also with w i. So, P w j will be 1 by n and P w i w j will also be 1 by n. So, if I put that here it will again come out to be this formula and; that means, it will the PMI will depend on w i if w i is frequent PMI will be low if it is rare it will become high I mean this is not desire.

To evaluate this problem, so they are 2 different things that you can do, one is you can probably start by the moving all the words that are having very very low occurrences. So, all the words that are occurring less than 2 or 3 times you can remove from your data all together and the words that are occurring more than that this problem will be not that bad, but if you want took also take into considerations the very very infrequent words. So, what you might have to do? You might have to takes into consideration some sort of bias.

(Refer Slide Time: 31:16)



PMI: Issues and Variations

Positive PMI

All PMI values less than zero are replaced with zero.

Bias towards infrequent events

Consider w_j having the maximum association with w_i,

$P_{corpus}(w_i) \approx P_{corpus}(w_j) \approx P_{corpus}(w_i, w_j)$

PMI increases as the probability of w_i decreases.

Also, consider a word w_j that occurs once in the corpus, also in the context of w_i. A discounting factor proposed by Pantel and Lin:

$$\delta_{ij} = \frac{f_{ij}}{f_{ij}+1}\frac{min(f_i, f_j)}{min(f_i, f_j)+1}$$

$PMI_{new}(w_i, w_j) = \delta_{ij} PMI(w_i, w_j)$

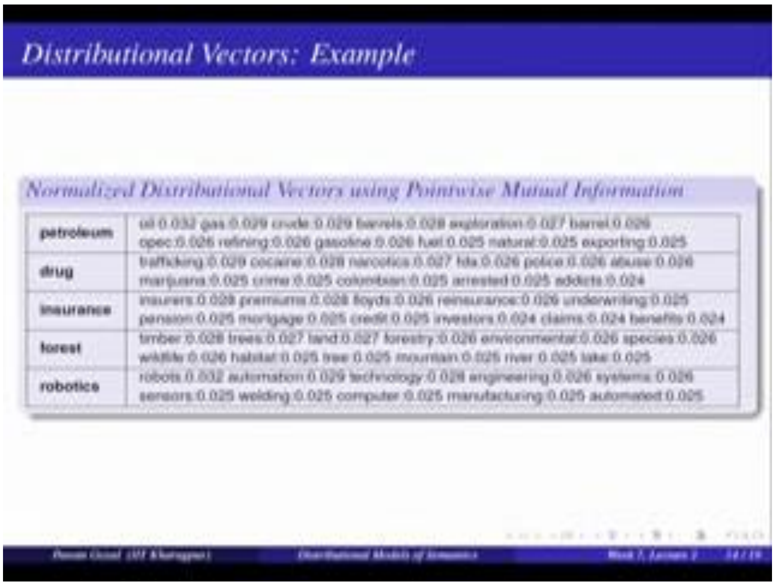
Pawan Goyal (IIT Kharagpur)
Distributional Models of Semantics
Week 7, Lec

.

What is the idea? There are these very important discounting factors that are proposed by Pantel and Lin. So, what it says that you multiply the PMI value with this discounting factor that is f i j divided by the f i j plus one times minimum of f i f j divided by minimum of f i f j plus 1. So, what would happen now? So, you will multiply this with PMI. So, is a frequency of the individual events are very very small this factor will become smaller. So, like that is take 2 cases here, one where f i is equal to f j is equal to f i j is 1, where they are even second f i is equal to f j is equal to f i j is equal to 10 not. So, rare, so in the first case what will be the discounting factors 1 divided by 1 plus 1 times 1 divided by 1 plus 1 this will be 1 by 4 in this case it will be 1 divided by 10 plus 1 times 1 divided by 10 plus 1 this will be sorry 10 divided by that. So, this will 100 divided by 121. So, discounting will be much more higher here. So, you will divide the PMI by 0.25, here you will or multiply by 0.25, here you multiply by roughly 0.8 and this takes care of the problems that we had talked about in with infrequent events.

Whenever there are infrequent PMI will become very high. So, if you use this discounting it will again come back to a reasonable value and this will not create a problem when the evens are not so rare.

(Refer Slide Time: 32:35)



Distributional Vectors: Example

Normalized Distributional Vectors using Pointwise Mutual Information

petroleum	oil:0.032 gas:0.029 crude:0.029 barrels:0.028 exploration:0.027 barrel:0.026 opec:0.026 refining:0.026 gasoline:0.026 fuel:0.025 natural:0.025 exporting:0.025
drug	trafficking:0.029 cocaine:0.028 narcotics:0.027 fda:0.026 police:0.026 abuse:0.026 marijuana:0.025 crime:0.025 colombian:0.025 arrested:0.025 addicts:0.024
insurance	insurers:0.028 premiums:0.028 lloyds:0.026 reinsurance:0.026 underwriting:0.025 pension:0.025 mortgage:0.025 credit:0.025 investors:0.024 claims:0.024 benefits:0.024
forest	timber:0.028 trees:0.027 land:0.027 forestry:0.026 environmental:0.026 species:0.026 wildlife:0.026 habitat:0.025 tree:0.025 mountain:0.025 river:0.025 lake:0.025
robotics	robots:0.032 automation:0.029 technology:0.028 engineering:0.026 systems:0.026 sensors:0.025 welding:0.025 computer:0.025 manufacturing:0.025 automated:0.025

Pawan Goyal (IIT Kharagpur)
Distributional Models of Semantics
Week 7, Lecture 2

Here is 1 example of what kind of vector do you get by taking this PMI. So, on the left hand side you have, so, what is done here? You taking a large corpus and finding out for individual words what are the other vectors that I having high PMI values and then you are doing some normalization. So, that all the PMI value is on this course for a given word add up to 1. So, what do you see here?

If I take a word like petroleum, you find words like oil, gas, crude, barrels, exploration that are coming out to be having very high PMI values, with drug you find words like trafficking, cocaine, narcotics, insurance, you find a words like insurers, premiums, Lloyds, with forest you find the words like timber, trees, land and robots; robotics you find word like robots, automation and so on and all this is coming out just by putting this PMI function over this corpus where different words and sentences are put together.

I am computing this function and finding out for a word what are similar words and you can see that it is capturing very nicely what are the other very similar words to that to s starting word like robotics you find words like robots and automation, but the word like forest you find immediately words like trees etcetera and this can give you nice intuition that you can use that very nicely to capture which 2 words are similar which 2 words are very very different and in some of application you can make use of that. So, I will start from here in the. So, for the next lecture and we will see how we can use that for some very interesting application like term (Refer Time: 34:38) information retrieval, I will define the problem and see how do you use that.

Thank you.