

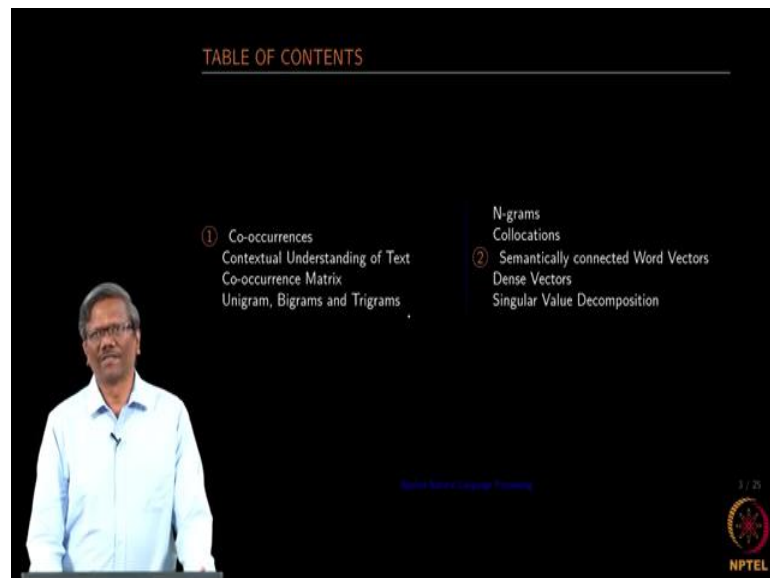
**Applied Natural Language Processing**  
**Prof. Ramaseshan Ramachandran**  
**Visiting Professor at Chennai Mathematical Institute**  
**Indian Institute Technology, Madras**

**Lecture - 14**  
**Contextual understanding of a text**

Hello everybody, in continuation of the discussion on words to vectors, today we will be diving a little deeper into a combination of words and terms and then finding out how the contexts of the surrounding words would help you and terms of understanding the meaning. So, we are going towards that path of slowly, from words finding the frequencies and then finding the inverse document frequency and then doing some kind of an empirical analysis on the corpus.

And finally, trying to in the last class trying to define, how we can convert the documents into a term, document matrix, and how we can convert the words into word vectors, document vectors, and how we can represent them in the vector space and so on ok.

(Refer Slide Time: 01:15)

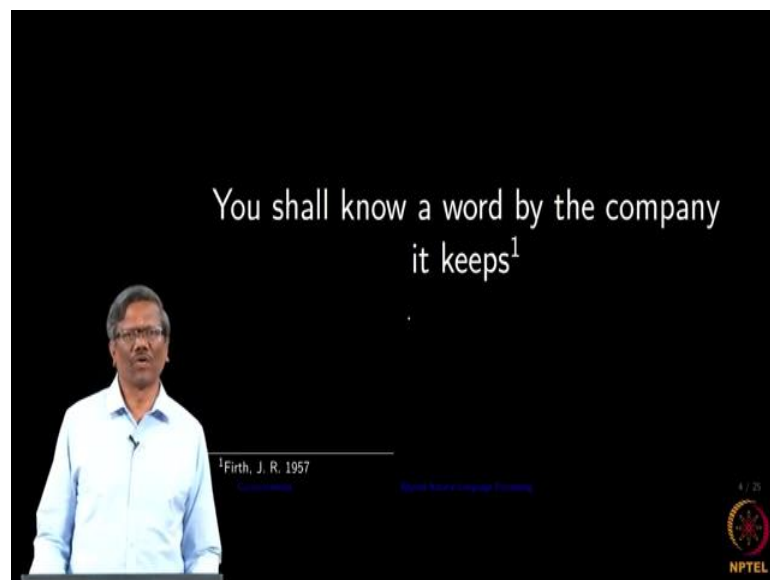


It is a continuation of that we going to be talking about the co-occurrences of words. You remember earlier we were only worried about a single word and trying to find out some statistics about that and then using the data that is available in the corpus, we were doing

some operations with respect to the corpus in terms of identifying the similar documents, extracting the name identities and so on so forthright.

In this class, we will be talking about the occurrences of words at together in a sentence and how we can exploit those occurrences and then we will see how those contextual understanding of the text would be done by using multiple words together, and then we will talk about how to create a co-occurrence matrix, we will talk about unigram, bigrams and trigrams or grams and then we will see how we can create semantically connected vectors using dense vectors formation, using one of the techniques from singular value decomposition.

(Refer Slide Time: 02:32)



Let us go to the next slide; this is a very important core that we need to remember. This is very profound with respect to the linguist analysis of the corpus as well as for all those who work on natural language processing.

In 1957 firth British linguists have made a profound statement saying that you shall know a word by the company it keeps. Meaning that it is very easy to understand the meaning of the word if we know what the context is right. So, it is possible for you to know some of the words that you do not know while reading a novel or a book, but if you go back and then read the sentence again, you probably would make up the meaning for a given word and then later going to the dictionary to find that it is almost similar to what you thought about that right. So, this is really a very very interesting and profound

statement that creates or a flurry of activities from the linguistic community as well as the linguist community as well as from the anal payee community.

(Refer Slide Time: 03:48)

The slide is titled "CONTEXTUAL UNDERSTANDING OF WORDS" in orange text at the top. Below the title, there is a list of five bullet points, each preceded by a red arrowhead. The speaker, a man with glasses wearing a light blue shirt, is visible in the lower-left corner of the slide. At the bottom of the slide, there is a citation: "Firth, J. R. 1957" and a small NPTEL logo in the bottom right corner.

- ▶ The study of *meaning* and *context* should be central to linguistics
- ▶ Exploiting the context-dependent nature of words
- ▶ Language patterns cannot be accounted for in terms of a single system
- ▶ The *collocation*, gives enough clue to understand a word and its meaning
- ▶ *No study of meaning apart from context can be taken seriously* <sup>2</sup>

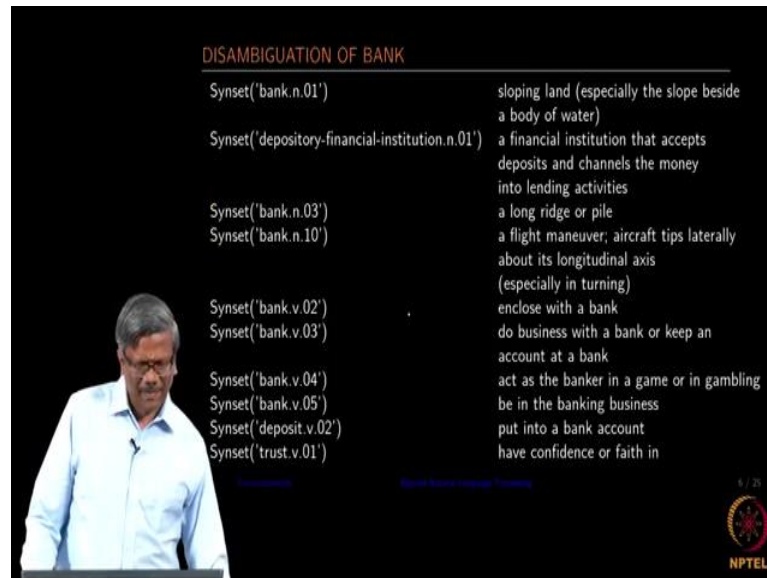
So, what this means? This means a study of the meaning and context should be central to linguistic, it's not possible for you to just know the word alone and then trying to make something out of that. So, you need to really understand the context so, it is easy for you to study the meaning of the word in a different contexts so, what and then we also want to see how those contexts can be exploited to it in terms of understanding the meaning of the word.

So, this central to the entire exercises of natural language processing I, if mentioned earlier as well, is to really understand the meaning of the word by looking at the large corpus and trying to figure out what could be the meaning or different meanings that could that a particular word may have ok. There could be some pattern that you will see in the corpus that also would help you in terms of understanding the phrases. The collocations give you enough clues to really understand the meaning of the word.

So, it is not possible for you to understand the meaning of the word by just looking only the looking at only the dictionary. So, you may have look at the context surrounded by the word and so on. So, the last part of that emphasizes that, know the study of meaning apart from context can be taken very seriously. So, again going back to what a firth

mentions you shall know a word by the company it keeps. So, please remember this throughout this lecture.

(Refer Slide Time: 05:37)



**DISAMBIGUATION OF BANK**

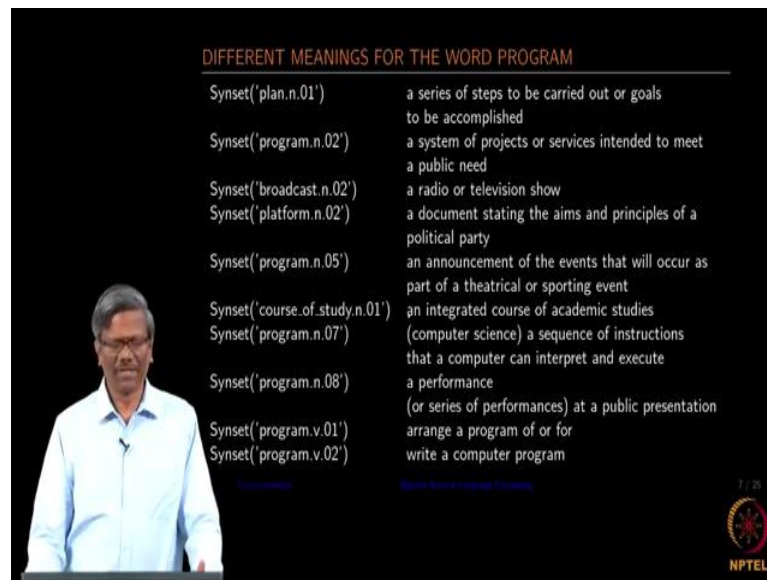
Synset('bank.n.01')	sloping land (especially the slope beside a body of water)
Synset('depository-financial-institution.n.01')	a financial institution that accepts deposits and channels the money into lending activities
Synset('bank.n.03')	a long ridge or pile
Synset('bank.n.10')	a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)
Synset('bank.v.02')	enclose with a bank
Synset('bank.v.03')	do business with a bank or keep an account at a bank
Synset('bank.v.04')	act as the banker in a game or in gambling
Synset('bank.v.05')	be in the banking business
Synset('deposit.v.02')	put into a bank account
Synset('trust.v.01')	have confidence or faith in

NPTEL

So, let us now look at the word bank right. So, bank means so, many different things if you go and then look at the thesaurus it will give you different flavors of the bank. So, one would be you know it is a sloping land which is by the side of the river right and then the second meaning could be related to the financial institute, the bank is the financial institution that accepts deposits and so on. And then if you go down you look at the last one, if you look at it you know it's you have some confidence or some faith in something right. So, you bank on something. So, that is another meaning.

So, if you look at the variations of the meaning of the bank, you know all could be derived by the context it is surrounded by. So, it is important for us to know where this word is used, in important you know what are all the words to the left and right of that bank as in order for you to really understand the meaning of that. Let us take one more example of the word program right.

(Refer Slide Time: 06:53)



DIFFERENT MEANINGS FOR THE WORD PROGRAM

Synset('plan.n.01')	a series of steps to be carried out or goals to be accomplished
Synset('program.n.02')	a system of projects or services intended to meet a public need
Synset('broadcast.n.02')	a radio or television show
Synset('platform.n.02')	a document stating the aims and principles of a political party
Synset('program.n.05')	an announcement of the events that will occur as part of a theatrical or sporting event
Synset('course.of.study.n.01')	an integrated course of academic studies
Synset('program.n.07')	(computer science) a sequence of instructions that a computer can interpret and execute
Synset('program.n.08')	a performance (or series of performances) at a public presentation
Synset('program.v.01')	arrange a program of or for
Synset('program.v.02')	write a computer program

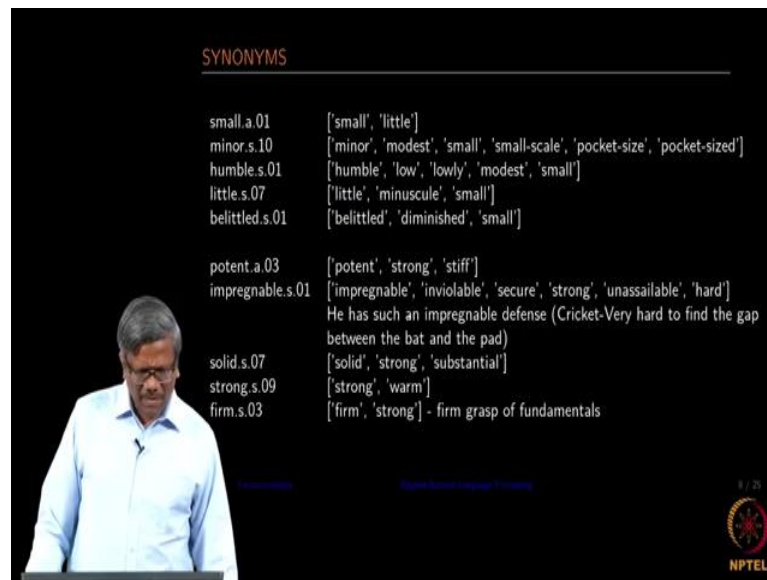
7 / 25  
NPTEL

So, if you were a computer science student, you will know that you know program means is a sequence of instructions that is the first thing that will come into your mind. Right, but that is not the only meaning that you have; you have several meanings you know it could be related to a system of projects or services indent to meet some public need ok. It could be a radio or a television show. So, it could be documents stating the aims and principles of a political party, it is a program of a political party correct.

So, the last one is what I mention you know write a computer program. All the meanings of the word program could be understood only when you understand the entire sentence or if you look at the context word surrounding the program, you should be able to understand the meaning of the word. So, it is very important for the data modal based approach in NLP to really know the context surrounding a word.

For you to translate you need to know the context surrounding that particular word. So, in order for you to generate a text or a sentence automatically, you need to understand the context in which this word is going to be used.

(Refer Slide Time: 08:23)



**SYNONYMS**

small.a.01	['small', 'little']
minor.s.10	['minor', 'modest', 'small', 'small-scale', 'pocket-size', 'pocket-sized']
humble.s.01	['humble', 'low', 'lowly', 'modest', 'small']
little.s.07	['little', 'minuscule', 'small']
belittled.s.01	['belittled', 'diminished', 'small']
potent.a.03	['potent', 'strong', 'stiff']
impregnable.s.01	['impregnable', 'inviolable', 'secure', 'strong', 'unassailable', 'hard']
	He has such an impregnable defense (Cricket-Very hard to find the gap between the bat and the pad)
solid.s.07	['solid', 'strong', 'substantial']
strong.s.09	['strong', 'warm']
firm.s.03	['firm', 'strong'] - firm grasp of fundamentals

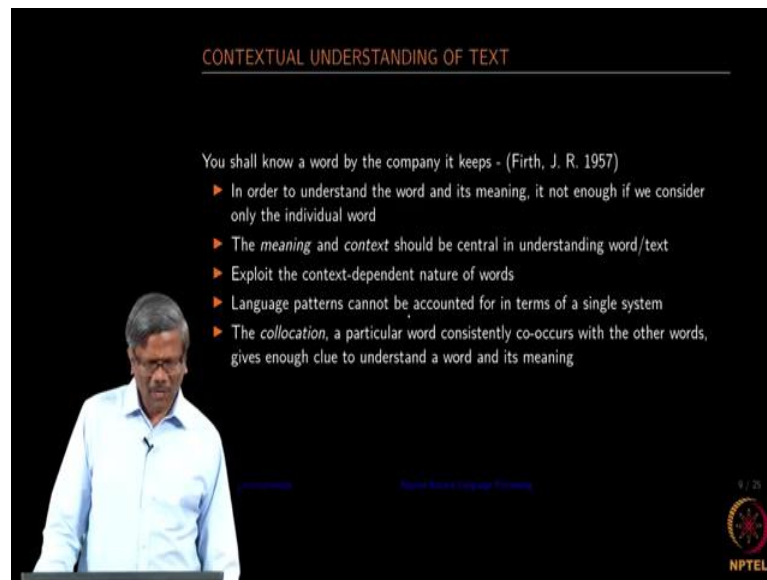
8 / 25

NPTEL

If you look at another example where we were listening only to these synonyms. For example, if you take the word small there are several synonyms also we hear at with that right. It is small could be little minor modest small small-scale and so on so forth. So, all mean the same thing. So, in order for us to really represent this particular synonym in our vectors space, we probably would like to use only one access rather than multiple access with respect to all the synonyms.

So, we really have to not only understand the context surrounding the word, but we also need to understand these synonyms a for the given word in order for us to really reduce the vector space on which we would like to operate on ok.

(Refer Slide Time: 09:21)



The slide features a black background with white text. At the top, the title 'CONTEXTUAL UNDERSTANDING OF TEXT' is displayed in a light blue font. Below the title, a quote is presented: 'You shall know a word by the company it keeps - (Firth, J. R. 1957)'. This is followed by four bullet points, each starting with a blue arrowhead. The first bullet point states that understanding a word's meaning is not enough if only the individual word is considered. The second emphasizes that meaning and context are central to understanding. The third notes that language patterns cannot be explained by a single system. The fourth defines collocation as a word that consistently co-occurs with others, providing clues to its meaning. In the bottom left, a man in a light blue shirt is shown from the chest up, appearing to be speaking. In the bottom right, there is a small red circular logo with a white design inside, and the text 'NPTEL' below it.

CONTEXTUAL UNDERSTANDING OF TEXT

You shall know a word by the company it keeps - (Firth, J. R. 1957)

- ▶ In order to understand the word and its meaning, it not enough if we consider only the individual word
- ▶ The *meaning* and *context* should be central in understanding word/text
- ▶ Exploit the context-dependent nature of words
- ▶ Language patterns cannot be accounted for in terms of a single system
- ▶ The *collocation*, a particular word consistently co-occurs with the other words, gives enough clue to understand a word and its meaning

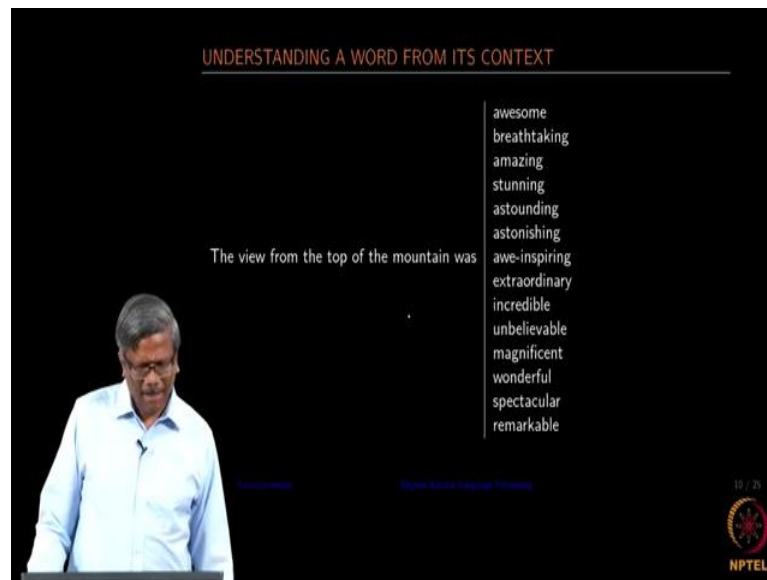
NPTEL

So, in order to understand the word and its meaning it is not enough if we consider only the individual word I as spoke about this at length earlier. At the meaning and context should be central to be understanding word or text right.

Exploit the context-dependent nature of the word. So, we need to really understand under which context these words are used and then similar context if the similar words would you similar context is another example that we need to remember or another statement that we need to remember. Language patterns cannot be accounted for in terms of a single system. So, we need to be able to look at various aspects in terms of understanding the patterns.

So, it is important for us to define these term collocations; that means, a particular word consistently co-occurs with the other word gives enough clue to understanding the word and its a meaning and this particular pattern does not occur by chance. So, it has happened several times. So, we should be able to really make some sense out of those some patterns and make use of that in our NLP applications.

(Refer Slide Time: 10:45)



To give me another flavor of this right it is you know how these synonyms could be used and found in the data-driven approach. Let us take one example sentence the view from the top of the mountain was awesome, breathtaking, amazing, stunning, astounding, astonishing, awe-inspiring, extraordinary, incredibly unbelievable, magnificent, wonderful, spectacular, remarkable right.

So, we can use any of these words in this context and remain the same is not it? So, suppose if we had a received a corpus that talks about the mountain, the valleys, the views from the top, the view of the valley from the top and so on and if that corpus was written by other authors, you will find that somewhat similar occurrences of this particular sentence by every author right.

Supposing if we were only looking at the positive aspects of the positive sentiments from the corpus or you would see that one author would have written that view from the top of the mountain was awe-inspiring; one author would have written that the view from the top of the mountain was remarkable; one could have written the view from the top of the mountain was breathtaking. So, now, you get the sense of this right.

So, this context that we were talking about is one the left side, and then the words are on the right sides. So, if you were using a similar context and there are several words there are used for this similar context; that means, we are going to be looking at the synonyms of a given word. So, when you look at this particular example, you should be able to pick



up of various of the words as similar words ah; that means, given similar context or a given a context and if the words occur in that particular context; that means, those words are synonyms.

This is one I think that we need to remember is the data-driven approach. So, this is not going to happen by a chance, it is going to happen by the count of occurrences of a given word with the given context. So, this is what we were going to be exploiting in our data-driven approach.