

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

Lecture - 31
Distributional Semantics – Introduction

Welcome everyone to the 7th week of the course. So, with that we are now moving into the second half. So, in the first 6 weeks we had talked about basic tasks on in any language, how do we pre process the language, how do we deal about its morphology and syntax. Now we will be going to some advance topics. So, in the next 3 weeks, starting this week, we were talking about semantics, how do we capture semantics?

Again semantics is very huge topic, we will focus only on very very basic and interesting notions that can you help you build some important applications and also read some research papers in this field. Later in the last 3 weeks of the course, we will be going to various applications. So, today we are starting our discussion on semantics. So, and we will be taking about 2 notions distribution semantics and then lexical semantics. Before going into that let us see what do I mean by semantics? So, in this lecture today we will give the introduction to the topic of distributional semantics.

(Refer Slide Time: 01:27)

The slide is titled "Introduction" in a blue header. Below the header, there is a light blue box containing the text "What is Semantics?". Below this box, the text "The study of meaning: Relation between symbols and their denotata." is displayed, followed by the example sentence "John told Mary that the train moved out of the station at 3 o'clock." At the bottom of the slide, there is a footer with the text "Pawan Goyal (IIT Kharagpur)", "Distributional Semantics - Introduction", "Week 7, Lecture 1", and "2 / 14".

As such when I say semantics, what comes to your mind? In general syntax talks about arrangement, how are that different words are arrange in the language in when while talk

about semantics, it immediately we move from arrangement to some sort of meaning domain. So, I want to find out, what is the meaning of different words, how do they combine together to form the meaning of the larger unit that can be the sentence; that is in general the main field.

If I have to give a definition, I can say that semantics is the study of meaning that is, what is the relation between symbols and what do they denote? So, here suppose I have the sentence, John told Mary that the train moved out of the station at 3 o'clock, it is a sentence in the natural language contains multiple words, we know how to tokenize them and how to find the morphological tag and may be also the syntax and different representations. Now what do I mean by saying that now I want to go to semantics? In semantics, I would like to know, what all these words themselves mean separately. So, what do I mean by John told Mary, etcetera. So, these are all symbols, now what do they denote? And when they combine together in a sentence, what meaning are they putting? So, this is as I said very very vast field, lot of different research has happened and is happening, we will focus on mainly the word meaning. So, how do we say, what is the meaning of a word, what are different notions? So, in general, so when I talk about distribution semantics or any other model that we will cover in this course, we are trying to find out, how in general computers can produce such semantic representations. So, machines can interpret some sort of semantics and that is why the idea of computational comes in here.

(Refer Slide Time: 03:36)

Computational Semantics

Computational Semantics
The study of how to automate the process of constructing and reasoning with meaning representations of natural language expressions.

Methods in Computational Semantics generally fall in two categories:

- **Formal Semantics:** Construction of precise mathematical models of the relations between expressions in a natural language and the world.
$$\text{John chases a bat} \rightarrow \exists x[\text{bat}(x) \wedge \text{chase}(\text{john}, x)]$$
- **Distributional Semantics:** The study of statistical patterns of human word usage to extract semantics.

Pawan Goyal (IIT Kharagpur) Distributional Semantics - Introduction Week 7, Lecture 1 3 / 14

Computational semantics in general is the study of how we can automate this process of building these semantic representations and also reasoning with them. So, when we talk about the methods, in general there are 2 different methods, one are based on formal semantics that is how do I construct various mathematical models that can tell me, what is the relation between various expressions in the language and also relate them to whatever there is there in the word.

For example, so, if you have heard about read predicate logic that is what is done in formal semantics. So, if I have a sentence like this, John chases a bat and you want to produce a mathematical structure that denotes the meaning of the sentence, so I will put them in very very logical form like I will say so, for startup, (Refer Time: 04:33) I will say there is an x where x is a bat and John chases the bat. So, John chase becomes a predicate and I will say John chases x and x have already defined as a bat and like that I will define, how do I convert my natural language expression to some sort of this logical form and then I will also have rules on how do I infer from this logical expressions? How do I reduce one expression into another 1 and then so on?

And this is again a field of formal semantics and that is something we will not cover in this course, what is something apart from formal semantics that we can see and there where the data, a lot of data that we have can help us and this is where the field of distribution semantics comes in that is can I study the statistical patterns of human word usage to extract semantics.

Can I see how humans are using different words in th language to find out what is their semantics? And this is what will be the topic of this week, what is the type of distributional semantics? How they capture that and how can I use that for certain meaningful applications? So, this field of distributional semantics is mainly built upon this hypothesis of this hypothesis of distribution, distributional hypothesis that is prevent from many many years, what is the distributional hypothesis? Let us see some famous quotes about this.

(Refer Slide Time: 06:09)

Distributional Hypothesis

Distributional Hypothesis: Basic Intuition

"The meaning of a word is its use in language." (Wittgenstein, 1953)

"You know a word by the company it keeps." (Firth, 1957)

→ Word meaning (whatever it might be) is reflected in linguistic distributions.

"Words that occur in the same contexts tend to have similar meanings." (Zellig Harris, 1968)

→ Semantically similar words tend to have similar distributional patterns.

Pawan Goyal (IIT Kharagpur) Distributional Semantics - Introduction Week 7, Lecture 1 4 / 14

As earlier as in 1953 Wittgenstein said that the meaning of a word is its usage in language. So, that is you can know, what is the meaning of a word if you see how it is being used in the language?

Along similar lines in 1957, Firth said you know the word by the company it keeps, now what do I mean by the company of a word? What are the other words it occurs within my corpus on my language and that tell me about the word? Now going 1 step further, so here so, what these different quotes are mentioning that the word meaning whatever it might be, I do not care about what exactly is the meaning, but whatever it is, it should be reflected in the linguistic distributions that is the way the word has been used in the language that will tell me the word meaning. So, now, Zellig Harris in 1968 that gave this famous code that took this idea to one step ahead that is words that occur in the same context tend to have similar meanings.

Now I can talk about 2 words having similar or different meanings, if I can somehow measure their context and capture that and I will say 2 words are having similar meanings, if the context in which they occur are very very similar. So, the idea here is 2 words, if they are semantically similar, they tend to have similar distribution patterns. So, now, that is what we will be doing in this week, how do we capture distribution patterns of words and use that to find out similarity across words.

(Refer Slide Time: 07:56)

Distributional Semantics: a linguistic perspective

"If linguistics is to deal with meaning, it can only do so through distributional analysis." (Zellig Harris)

"If we consider words or morphemes A and B to be more different in meaning than A and C, then we will often find that the distributions of A and B are more different than the distributions of A and C. In other words, difference in meaning correlates with difference of distribution." (Zellig Harris, "Distributional Structure")

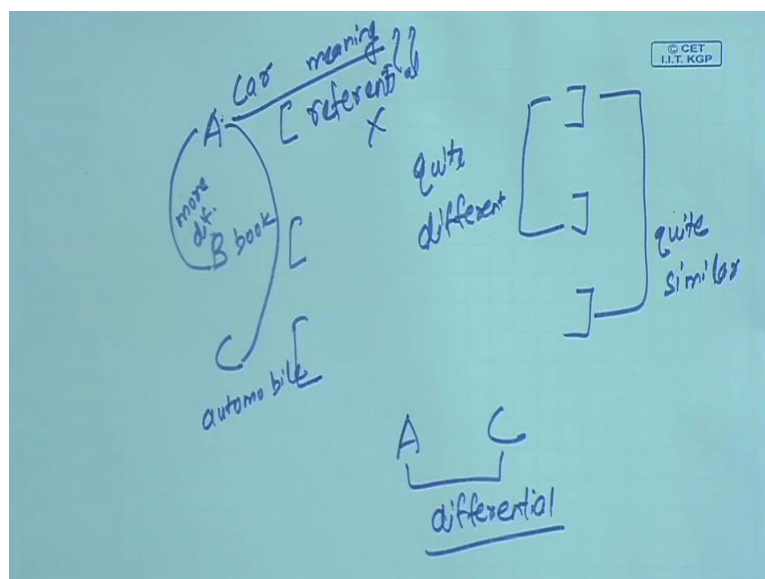
Differential and not *referential*

Pawan Goyal (IIT Kharagpur) Distributional Semantics - Introduction Week 7, Lecture 1 5 / 14

Let us see some other quotes from Zellig Harris. So, he also said that if linguist have to deal with meaning it can only do so through distributional analysis and that he said in very early 70s and 80s and if you see the (Refer Time: 08:12) field right now, so whatever the main methods or semantics are there, they are built upon distributions. So, with the idea of; so with the recent idea of word and biddings and all and this is what we will also capture in our in this week of this course.

Another very similar sort of code from Zellig Harris from his distributional structures, so if we consider, what are morphemes A and B to be more different in meaning than A and C then we will often find that the distributions of A and B are more different than the distributions of A and complaint in other words difference in meaning correlates with difference of distribution.

(Refer Slide Time: 08:55)



What he saying suppose I take 3 different words or morphemes A B and C and suppose there is some where I can capture the distribution. So, that is how they are occurring in the languages. So, I have distribution for A B and C, now what he is saying? If A and B, the distributions are more different than that of A and C then what you will find? So, what we are saying A and B are more different than A and C, if the distributions are more different in the meaning also, you will find they are more different. So, you can think of these as so, I take some examples like car as A and B C can be automobile, they are quite close and B can be something like a book. So, here a car and automobile are quite similar and car and book are more different. So, they are very very different. So, if this is what I see in meaning, I will also see something similar in distributions. So, what I will find? These 2 distributions are quite similar and these 2 would be quite different.

We will see in this week that how do we capture these distributions and how do we compare that, but this is the basic idea, now what is another important thing here? So, distributions that we are capturing, whatever semantics it is allowing me to handle, it is all differential not referential, now what is the difference between that 2 terms differential and referential? So, again if I would look at the same example of these 3 words; car, book and automobile, so what do I mean by saying so? The semantic sign capturing is only differential, but referential. So, I cannot say, what is the meaning of car? I am not defining the meaning of car at sort of this concept representation; there is something I am not doing. So, what I am doing? What, How similar or how different the

2 meanings are? So, I am saying A and C, how similar their meanings are or how different their meanings are? That is why this is differential sort of understanding of semantics. So, how different or how similar 2 different meanings are I am not talking about some referential meaning in distributional semantics.

Now if we look at so this was the linguistic prospective is there some cognitive prospective also with distributional semantics.

(Refer Slide Time: 11:52)

Distributional Semantics: a cognitive perspective

Contextual representation
A word's contextual representation is an abstract cognitive structure that accumulates from encounters with the word in various linguistic contexts.

We learn new words based on contextual cues
He filled the **wampimuk** with the substance, passed it around and we all drunk some.
We found a little **wampimuk** sleeping behind the tree.

Pawan Goyal (IIT Kharagpur) Distributional Semantics - Introduction Week 7, Lec. 14

There we have this idea that what is the representation of the word and it is said to be some sort of abstract cognible structure that I am storing in my brain or somewhere we do not know, but that is something I gather as I keep on hearing this word or looking or finding this word in more and more context as I find this word in more and more context, I tend to build some sort of representation about this word this is the cognitive prospective and what is some sort of evidence. So, for example, when you encounter a new word that you have never heard before even if we do not know its meaning you may guess something about this word what it might be?

Let us take an example like I am saying this word wampimuk, I have never heard this word before, now I am seeing this word or hearing this word in this context, he filled the wampimuk with the substance, passed it around and we all drunk some. So, just by looking at the word that are occurring around this word and seeing what are the words might occur in place of wampimuk, I can say that this might be some sort of a container

some sort of a glass or something so which can be used for filling up the substances and passing around.

As I keep on hearing a word or seeing a word in more and more context, I tend to build some sort of meaning structure about that word. Now suppose I have heard this word in a very different context like this, we find a little wampimuk sleeping behind the tree, immediately, I will have a different interpretation, I will say may be wampimuk some is some sort of a small animal. So, this is some sort of intuition behind the cognitive prospective and why we might think the meaning in terms of the distributions in the language.

(Refer Slide Time: 13:47)

Distributional Semantic Models (DSMs)

- Computational models that build contextual semantic representations from corpus data
- DSMs are models for semantic representations
 - ▶ The semantic content is represented by a vector
 - ▶ Vectors are obtained through the statistical analysis of the linguistic contexts of a word
- Alternative names
 - ▶ corpus-based semantics
 - ▶ statistical semantics
 - ▶ geometrical models of meaning
 - ▶ vector semantics
 - ▶ word space models

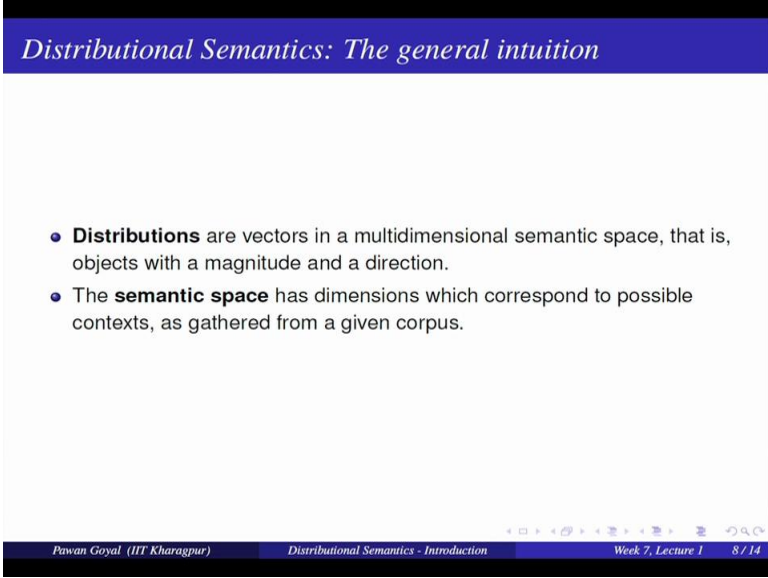
Pawan Goyal (IIT Kharagpur) Distributional Semantics - Introduction Week 7, Lecture 1 7 / 14

Now, to capture these distribution and then and find semantics from there, the models that are used are called distributional semantic models and also there is some term called DSM for them and what are these? So, these are various computational models that build contextual semantic representations from my corpus data. So, now, what is important here is that I am already given some sort of and if more data generally the better. So, I am given some corpus data on how different words are used in language and from there I am trying to come up with some semantic models distribution semantic model and this will capture the differential aspects of meaning among words.

DSMs are models for semantic representation and where I capture a semantic content by using a vector. So, for each word we will try to build a different vector that will capture

the semantics and these vectors are obtained via the statistics analysis of the linguistic contexts of the word. Now this word occurs in various context will help to find out its vector representation and there are some alternative names for this semantics like corpus based semantics statistical semantics geometrical models of meaning vector semantics and word space models and there might be some other prevalent names also, but we capture the same idea that can I use the distribution of the words to find out the meaning or at least capture which words are similar than other which pair of words are similar than other pair of words.

(Refer Slide Time: 15:26)



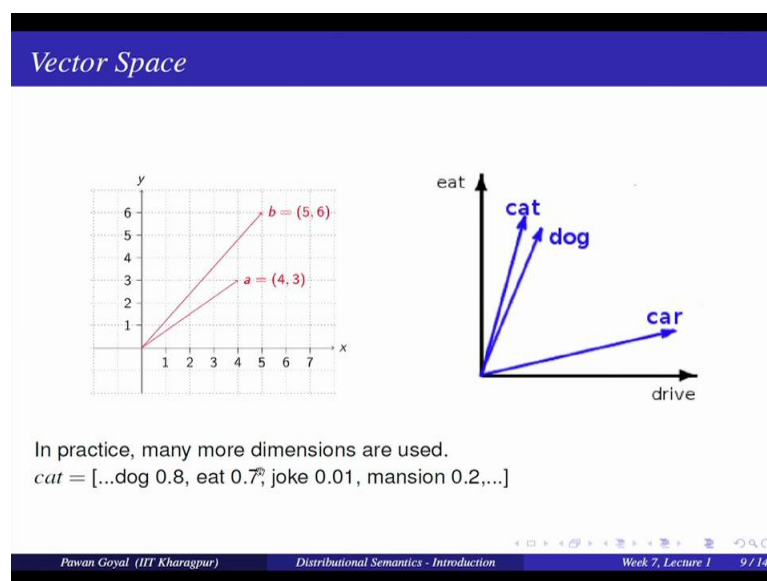
Distributional Semantics: The general intuition

- **Distributions** are vectors in a multidimensional semantic space, that is, objects with a magnitude and a direction.
- The **semantic space** has dimensions which correspond to possible contexts, as gathered from a given corpus.

Pawan Goyal (IIT Kharagpur) Distributional Semantics - Introduction Week 7, Lecture 1 8 / 14

Now again so, this distribution semantics when I talk about, what do you mean by the 2 terms distribution semantics? So, when I say distribution, they are the vectors denoting different different words and their vectors in some multi high dimensional or in low dimensional space and this space is the semantic space. So, I have distributions that are vectors denoting the words in that multidimensional semantic space and semantics space has dimensions which correspond to various possible contexts and this context can be gathered from a corpus. So, what I am doing? Every word I am denoting in this semantics space and this is my distribution semantics and this semantics space composed of various dimensions that are my context in which I am trying to represent a given word. So, let us take this an example.

(Refer Slide Time: 16:27)



When I talk about symbol vector space model, so that all of you would already be aware of, so this is my 2 dimensional plane and it is x y plane and I can denote different objects. So, here there are 2 points that I am denoting a and b by their coordinates. So, what are, so their projection in on x axis and y axis that tells me the coordinates and I can now capture how similar, how near to words 2 objects are in this space analogous to that, now let us think of a semantics space where dimensions are not x and y axis, but they are various context and can I denote all my words in those dimensions.

So, here suppose that my dimensions are 2 words eat and drive. 2 different words and I am trying to denote 3 words; car, dog and cat in these dimensions and what you are seen here? So, this might denote how often a word occurs with this word on and this word the projection will denote that. So, you see cat comes quite often with the word eat similarly dog car comes a lot often with drive, but not so much with eat. So, you see immediately when you try to put these word in this semantic space will capture some similarity that cat and dog are probably similar much more similar than cat and car and dog and car and this is the idea, can I use different context as my dimensions and represents all my words in those dimensions to give a meaning representation? In general here, I have shown you only for the dimension that corresponds to 2 words, but in general you all have to use any number of dimensions.

So, you can use any number of words that you want. So, so you might have a representation like that. So, I can say cat is a any other object or a word that has a weight of point eight in the dimension of dog 0.7 dimension of eat 0.01 dimension of joke. So, depending on how often the word cat occurs with all these words and this I can do for different different words and then I can capture how similar they are and this is a very very powerful technique although its looks very simple it works quite well in many many applications that that we will see.

(Refer Slide Time: 18:49)

Word Space

Small Dataset

An automobile is a wheeled motor vehicle used for transporting passengers .
A car is a form of transport , usually with four wheels and the capacity to carry around five passengers .
Transport for the London games is limited , with spectators strongly advised to avoid the use of cars .
The London 2012 soccer tournament began yesterday , with plenty of goals in the opening matches .
Giggs scored the first goal of the football tournament at Wembley , North London .
Bellamy was largely a passenger in the football match , playing no part in either goal .

Target words: {automobile, car, soccer, football}

Term vocabulary: {wheel, transport, passenger, tournament, London, goal, match}

Pawan Goyal (IIT Kharagpur) Distributional Semantics - Introduction Week 7, Lecture 1 10 / 14

Now let us take an example and how do we start constructing this a vector space or word space model and try to see can we compute the similarity across words. So, here I am given a small data set that has 6 different sentences in. So, like an automobile is a wheeled motor vehicle used for transporting passengers and so on, there are 6 different sentences are given and I have to build a vector space model or a distributional semantic model now to build the model, I need to start with what are the words I want to represent and these are my target words. So, here I want to represent 4 word automobile car soccer and football. So, now, once I know, what are the words I want to represent? The next question I need to ask is I need to represent in what dimensions? So, this becomes my context, here also denoted by term vocabulary what are the dimensions in which I will be denoting all these 4 words. So, here I have 1, 2, 3, 4, 5, 6, 7, dimensions.

(Refer Slide Time: 19:54)

Constructing Word spaces

Informal algorithm for constructing word spaces

- Pick the words you are interested in: **target words**
- Define a **context window**, number of words surrounding target word
 - ▶ The context can in general be defined in terms of documents, paragraphs or sentences.
- Count number of times the target word co-occurs with the context words:
co-occurrence matrix
- Build vectors out of (a function of) these co-occurrence counts

Pawan Goyal (IIT Kharagpur) Distributional Semantics - Introduction Week 7, Lecture 1 11 / 14

Now so what will be the informal algorithm for constructing the distributional semantic space or the word space, you start by picking up the words that you are interested in; that means, the words for which you want to find the distributions and these are called your target words. So, here I start with 4 target words then you define, what are the context words also? So, here you found 7 context words, next question would be when do I say that this word occurs with another word should I say that when it occurs in the same sentence same paragraph same document or within plus minus 2 words when do I say that. So, that is called the context window.

So, I can define a context window to be may be 5 words around the word or it can be a sentence paragraph and whatever word. So, I need to define a context window that is how many words should I consider surrounding a target word and it can be defined in terms of a document paragraph sentences and so on.

Now, a simple method would be once you defined the context window you know, what are your target words what are your context words? So, when you are defining a distributions of the target words find out how often they occur with the various context words within the context window and just write down the number this occurs with this word in this context window 2 times 3 times 0 times and so on and you define the vectors for different target words and this is this can also be called as your co occurrence matrix between the target words and the context words now once you have built the co

occurrence matrix you can think of each row of this matrix as your vector for that denotes this individual word.

(Refer Slide Time: 21:48)

Computing similarity							
	wheel	transport	passenger	tournament	London	goal	match
automobile	1	1	1	0	0	0	0
car	1	2	1	0	1	0	0
soccer	0	0	0	1	1	1	1
football	0	0	1	1	1	2	1

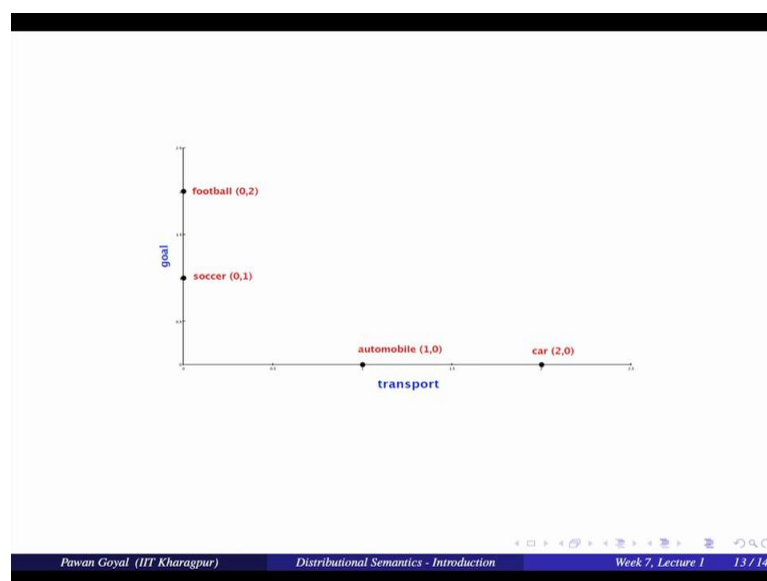
Using simple vector product

automobile . car = 4	car . soccer = 1
automobile . soccer = 0	car . football = 2
automobile . football = 1	soccer . football = 5

If we try to apply this algorithm on the previous example that we have seen, so what we will have? We will have 4 words as the target words, 7 word as my context words and I will have a matrix of dimension 4 cross 7 and each element will denote how often this target word occurred with this context word within my context window.

Let us see the whole sentence here in the context window. So, you can try doing that with the example that that we had seen and try to find out what are the different number of times each individual word occurs with another word and you will find something like that. So, this you can call as your co occurrence matrix or distribution matrix composed of targets and contexts. So, what do we see here? Automobile occurs with wheel, transport, passenger, football, occurs with passenger, tournament, London, goal and match and similarly for car and soccer. So, now, this is my representation of target words in this semantic space.

(Refer Slide Time: 22:53)



Now, suppose I take only 2 dimensions of this semantic space. So, here I had 1, 2, 3, 4, 5, 6, 7, dimensions; suppose I see only 2 dimensions. So, here I am seeing goal and transport and I am projecting all the 4 words in this dimensions. So, what representation I will see? So, I will see that soccer occurs with goal and goal not be transport, football occurs with goal not be transport, automobile occurs with transport not with goal and car occurs with transport not with goal. Immediately you can see some separation that is soccer and football are coming out with more similar automobile and car are coming out with excuse me are being more similar by just looking at these 2 dimensions and that is some very interesting aspect that the distributional semantic models try to capture.

Now, suppose I go back to my vector representation, so, where I have all the 7 dimensions and I want to find out which 2 words are similar and to what degree. So, one simple thing I can do is to take the dot products. So, that will tell me if they are having weights in similar dimensions if they having weights in different dimensions the dot products will be 0 or close to 0, but if they are having weights in same dimensions their dot product will be high suppose we are not doing any normalization, I can just take simple dot products and find out which 2 objects or which 2 words are more similar than other words. So, if you try to do that in this example, so what did you would you find? Automobile and car, so dot product would be 1 plus 2 3 plus 1 4 and everything else is 0.

Similarly, automobile soccer you will find 0, automobile football 1, car soccer 1, car football 2 and soccer football 5. So, from these simple dot products computations, so what is something that you see? So, you see that immediately you can capture the idea that here automobile and car are very similar. They are having a very high dot product 4, similarly soccer and football are also very similar, they are having a dot product of 5, on the other hand, automobile soccer are not very not much similar car and soccer are not much similar and so on. So, you can find out some sort of differential aspect of my meaning that is which 2 words have similar meanings and which 2 words are different meanings just by looking at the simple distribution and as you keep on getting more and more data you will you will be able to capture more and more information about how similar and how different 2 meanings of words are this is a very very basic idea that that I had tried to give in this lecture.

Now we will see, what are the different formal ways in which we can construct these models, how do we count what are different ways in which we can denote the entries of this matrix and how do we use that to compute similarity across 2 different words. So, this we will cover in more detail in the next lecture, but I hope that the main idea of using these distribution models is clear.

Thank you.