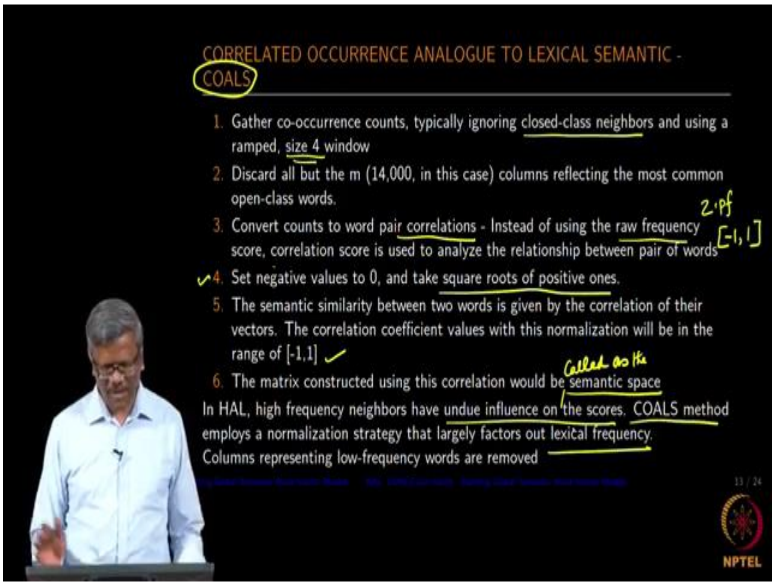
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture – 90
Correlated Occurrence Analogue to Lexical Semantic - COALS

(Refer Slide Time: 00:15)



CORRELATED OCCURRENCE ANALOGUE TO LEXICAL SEMANTIC - COALS

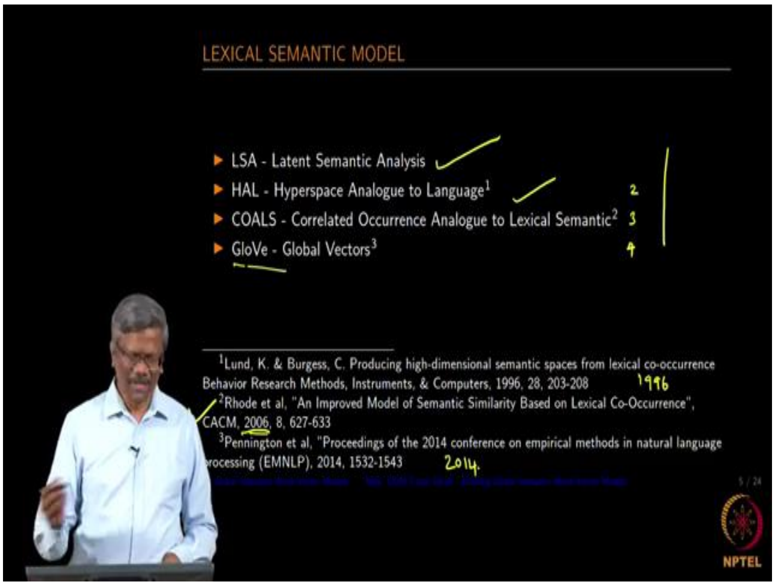
1. Gather co-occurrence counts, typically ignoring closed-class neighbors and using a ramped, size 4 window
2. Discard all but the m (14,000, in this case) columns reflecting the most common open-class words.
3. Convert counts to word pair correlations - Instead of using the raw frequency score, correlation score is used to analyze the relationship between pair of words 2·pf [-1, 1]
4. ✓ Set negative values to 0, and take square roots of positive ones.
5. The semantic similarity between two words is given by the correlation of their vectors. The correlation coefficient values with this normalization will be in the range of [-1,1] ✓
6. The matrix constructed using this correlation would be called as the semantic space

In HAL, high frequency neighbors have undue influence on the scores. COALS method employs a normalization strategy that largely factors out lexical frequency. Columns representing low-frequency words are removed

13 / 24

So, this is called as the Correlated Occurrences Analogue to Lexical Semantic. So, that is the name given by the authors it is the acronymous COALS. And, let us see what these authors have done.

(Refer Slide Time: 00:40)



LEXICAL SEMANTIC MODEL

- LSA - Latent Semantic Analysis
- HAL - Hyperspace Analogue to Language[1] 2
- COALS - Correlated Occurrence Analogue to Lexical Semantic[2] 3
- GloVe - Global Vectors[3] 4

[1] Lund, K. & Burgess, C. Producing high-dimensional semantic spaces from lexical co-occurrence Behavior Research Methods, Instruments, & Computers, 1996, 28, 203-208 1996

[2] Rhode et al, "An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence", CACM, 2006, 8, 627-633

[3] Pennington et al, "Proceedings of the 2014 conference on empirical methods in natural language rocessing (EMNLP), 2014, 1532-1543 2014.

5 / 24

So, again this paper is available you should read again this is a very simple paper to read; this one. An Improved Model of Semantics Similarity Based on Lexical Co-Occurrence by Rhode and others ok. So, this was published in 2006 CACM. They are not very far away ok. So, for the first one, I think it was around 1996, this paper is in 2006 and the third one that we will talk about is in 2014 again this is not a supervised model, the third one also is not a supervised model, ok.

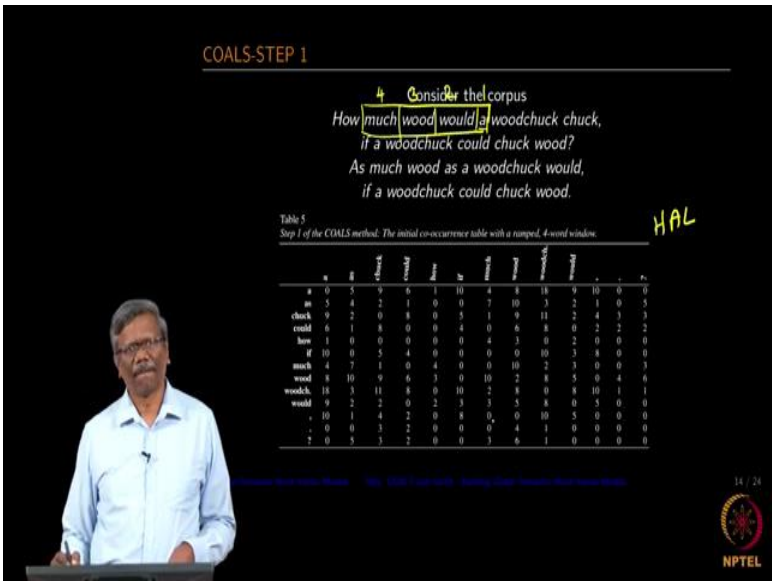
So, in this case, again it is very similar to have. We going be capture in the co-occurrence counts. Typically ignoring closed class neighbors. Let us find out what they are little later using a ramped window whose size is 4. So, it is very similar to HAL, but in this case, we are not using a ramped window of size 10. But we are using a ramped window of 4. Discard all but, the m columns reflecting the most common open class, right. The high-frequency words are taken off. We are taking only the most common words.

Convert the counts to word pair correlations. So, this is one change that we are making here. We are introducing the correlations here instead of just the counts. So, instead of using the raw frequency score; the correlation score is used to analyze the relationship between the pair of words, ok. The mechanism of building the term matrix is the same. The only thing is the window size is 4 and we still have the ramped window. And then in this case, we are using a correlation where the values would be in the range of minus 1 to plus 1 ok.

So, that is what the fourth bullet point is saying. Set the negative values to 0 and take these square roots of the positive 1. This is somewhat odd there could be different ways of doing this. Because, the values of the correlations are very small they want to scale it up and they use a square root or option. The semantic similarity between two words is given by the correlation of their vectors, ok. The correlation coefficient values with this normalization will be in the range of minus 1 to plus 1.

The matrix constructed using this correlation would be called as the semantic space, right. Since, we are creating the word vectors which are actually bringing out similar words in the same vector space we call it as this semantics right. In HAL high-frequency neighbors have undue influence on the scores. In COAL we use a normalization strategy that, largely factors out lexical frequency. Columns representing low-frequency words are also removed.

(Refer Slide Time: 04:22)



COALS-STEP 1

Consider the corpus

How much wood would a woodchuck chuck,
if a woodchuck could chuck wood?
As much wood as a woodchuck would,
if a woodchuck could chuck wood.

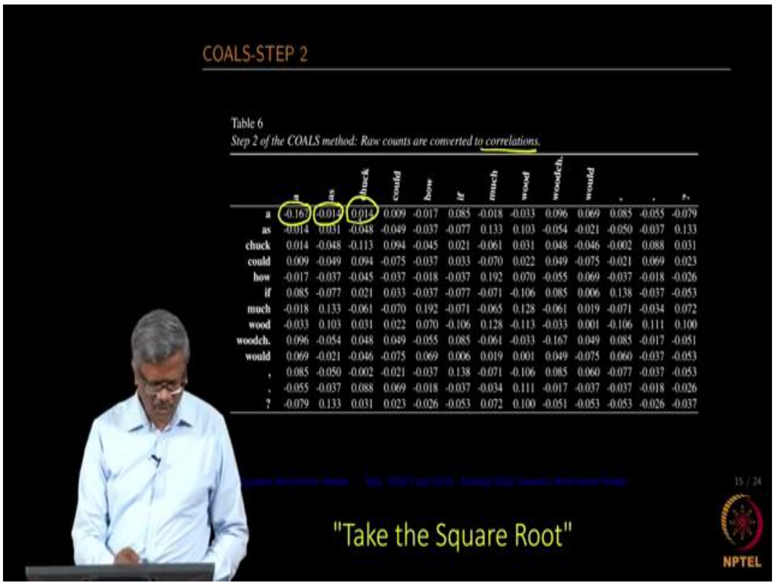
Table 5
Step 1 of the COALS method: The initial co-occurrence table with a ramped, 4-word window.

	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	0	5	9	6	1	10	4	8	18	9	10	0	0
as	5	4	2	1	0	0	7	10	3	2	1	0	5
chuck	9	2	0	8	0	5	1	9	11	2	4	3	3
could	6	1	8	0	0	4	0	6	8	0	2	2	2
how	1	0	0	0	0	0	4	3	0	2	0	0	0
if	10	0	5	4	0	0	0	0	10	3	8	0	0
much	4	7	1	0	4	0	0	10	2	3	0	0	3
wood	8	10	9	6	3	0	10	2	8	5	0	4	6
woodch.	18	3	11	8	0	10	2	8	0	8	10	1	1
would	9	2	2	0	2	3	3	5	8	0	5	0	0
,	10	1	4	2	0	8	0	0	10	5	0	0	0
.	0	0	3	2	0	0	0	4	1	0	0	0	0
?	0	5	3	2	0	0	3	6	1	0	0	0	0

So, they are considered as noise. So, they are removed. High-frequency ones are removed, low frequency ones are also removed ok. So, let us see how this could be done again. I do not want to get into the details of this, if you want to build this you start in this fashion right. For now, you are going to have 4 3, I am sorry 2 and 1. And then you keep moving that a ramped window over the entire corpus and then start computing the count of the co-occurrences.

So, the first step is to create the co-occurrence matrix, it is very similar to HAL. There is no are different except that we do not do from right to left. There is only one in one direction we are doing. In HAL it is direction sensitive, whereas, in this case it is one direction. So, if you look at this particular matrix built using this you will see asymmetry ok.

(Refer Slide Time: 05:27)



COALS-STEP 2

Table 6
Step 2 of the COALS method: Raw counts are converted to correlations.

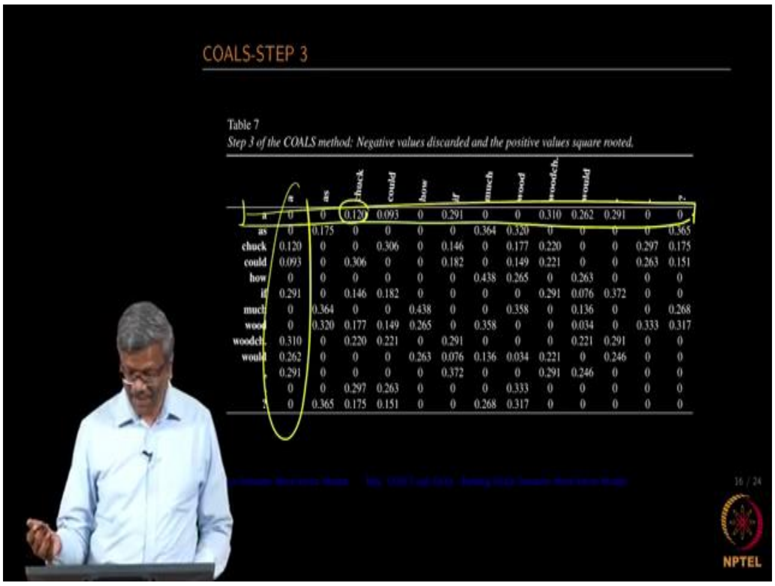
	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	-0.167	-0.014	0.014	0.009	-0.017	0.085	-0.018	-0.033	0.096	0.069	0.085	-0.055	-0.079
as	-0.014	0.031	-0.048	-0.049	-0.037	-0.077	0.133	0.103	-0.054	-0.021	-0.050	-0.037	0.133
chuck	0.014	-0.048	-0.113	0.094	-0.045	0.021	-0.061	0.031	0.048	-0.046	-0.002	0.088	0.031
could	0.009	-0.049	0.094	-0.075	-0.037	0.033	-0.070	0.022	0.049	-0.075	-0.021	0.069	0.023
how	-0.017	-0.037	-0.045	-0.037	-0.018	-0.037	0.192	0.070	-0.055	0.069	-0.037	-0.018	-0.026
if	0.085	-0.077	0.021	0.033	-0.037	-0.077	-0.071	-0.106	0.085	0.006	0.138	-0.037	-0.053
much	-0.018	0.133	-0.061	-0.070	0.192	-0.071	-0.065	0.128	-0.061	0.019	-0.071	-0.034	0.072
wood	-0.033	0.103	0.031	0.022	0.070	-0.106	0.128	-0.113	-0.033	0.001	-0.106	0.111	0.100
woodch.	0.096	-0.054	0.048	0.049	-0.055	0.085	-0.061	-0.033	-0.167	0.049	0.085	-0.017	-0.051
would	0.069	-0.021	-0.046	-0.075	0.069	0.006	0.019	0.001	0.049	-0.075	0.060	-0.037	-0.053
,	0.085	-0.050	-0.002	-0.021	-0.037	0.138	-0.071	-0.106	0.085	0.060	-0.077	-0.037	-0.053
.	-0.055	-0.037	0.088	0.069	-0.018	-0.037	-0.034	0.111	-0.017	-0.037	-0.037	-0.018	-0.026
?	-0.079	0.133	0.031	0.023	-0.026	-0.053	0.072	0.100	-0.051	-0.053	-0.053	-0.026	-0.037

15 / 24

"Take the Square Root"

And then in the second step, what we do is; we convert that into correlations I am sure you know how to compute the correlation between two random variables. So, in the same fashion we can compute the correlation between two words and then the values are negative and positives.

(Refer Slide Time: 05:52)



COALS-STEP 3

Table 7
Step 3 of the COALS method: Negative values discarded and the positive values square rooted.

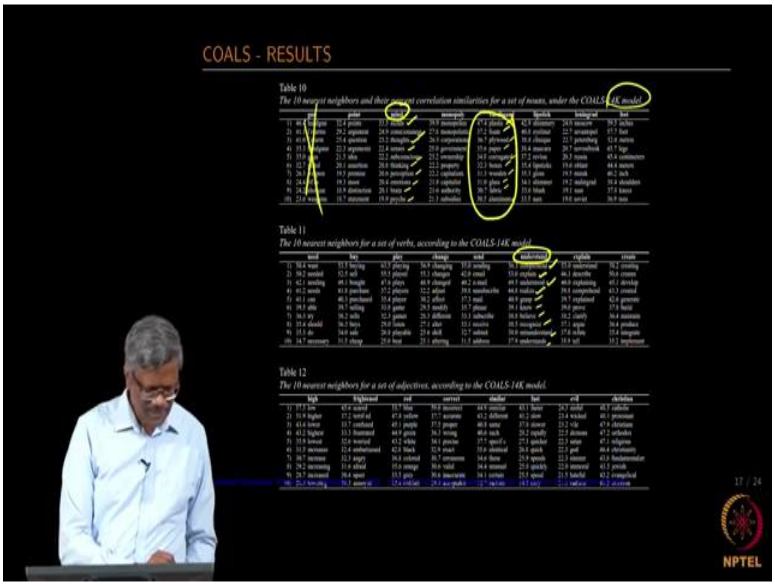
	a	as	chuck	could	how	if	much	wood	woodch.	would	,	.	?
a	0	0	0.120	0.093	0	0.291	0	0	0.310	0.262	0.291	0	0
as	0	0.175	0	0	0	0	0.364	0.320	0	0	0	0	0.365
chuck	0.120	0	0	0.306	0	0.146	0	0.177	0.220	0	0	0.297	0.175
could	0.093	0	0.306	0	0	0.182	0	0.149	0.221	0	0	0.263	0.151
how	0	0	0	0	0	0	0.438	0.265	0	0.263	0	0	0
if	0.291	0	0.146	0.182	0	0	0	0	0.291	0.076	0.372	0	0
much	0	0.364	0	0	0.438	0	0	0.358	0	0.136	0	0	0.268
wood	0	0.320	0.177	0.149	0.265	0	0.358	0	0	0.034	0	0.333	0.317
woodch.	0.310	0	0.220	0.221	0	0.291	0	0	0	0.221	0.291	0	0
would	0.262	0	0	0	0.263	0.076	0.136	0.034	0.221	0	0.246	0	0
,	0.291	0	0	0	0	0.372	0	0	0.291	0.246	0	0	0
.	0	0	0.297	0.263	0	0	0	0.333	0	0	0	0	0
?	0	0.365	0.175	0.151	0	0	0.268	0.317	0	0	0	0	0

16 / 24

If the values are all negative; what you are going to be doing is we are going to be taking off those negative values. And then, the authors are advising that; we should square the small values that you find here. For example: in the case of chuck, ok the value is 0,014. And they square it gets the rather square rooted the values are positive values are squarely rooted get this value ok.

So, this is step 3 part of it. And, then once step 3 is completed you have your word vectors. You can either take this one or this one. The row one right, they are going to be the same.

(Refer Slide Time: 06:43)



COALS - RESULTS

Table 10

The 10 nearest neighbors and their percent correlation similarities for a set of nouns, under the COALS-14K model

Table 11

The 10 nearest neighbors for a set of verbs, according to the COALS-14K model

Table 12

The 10 nearest neighbors for a set of adjectives, according to the COALS-14K model.

17 / 24

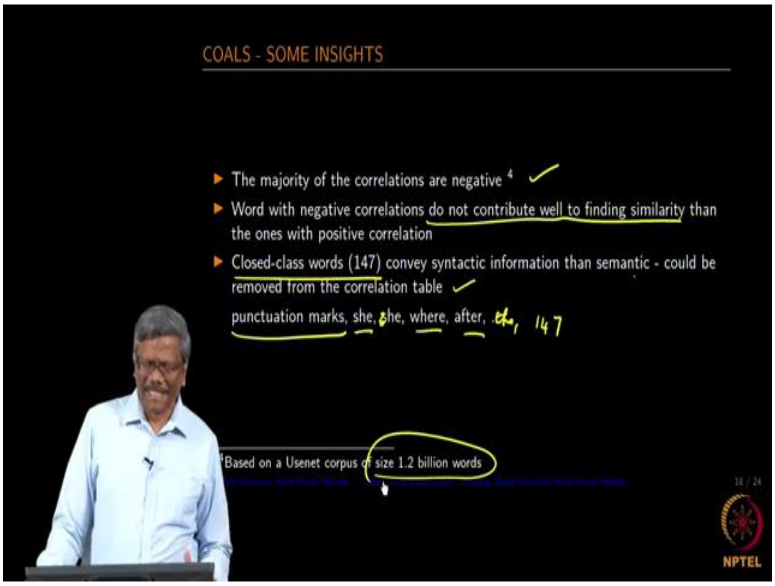
So, what they have done is; again, I am not going to go into the details of what happens when you do the correlation and so on. So, I am only going in and take the results that and then show, how this one is performing and then is it better than HAL and so on. So, in this case, again they have taken a set of nouns using a 14 k model. Because, they are taken just about 14 k columns and then trying to find out the similarity for a set of nouns here. So, they are going to be sorry I do not want to look at this, let us look at the mind, ok. So, if you look at the mind and the similar work that our COALS found was minds, consciousness, thought, senses, subconscious, thinking, perception, emotions, brain and psyche. I think it has done a good job here, ok.

Let us look at the cardboard, plastic foam, plywood paper, corrugated boxes, wooden glass, fabric aluminum. You remember, we saw the cardboard example of HAL. We had some variation there whereas, here you can find that the material which is very close or similar to cardboard is listed in this. So, by just looking at this list itself we can say COALS is producing better results than HAL. The reason could be due to the high removal of the high frequency and the low-frequency noises that you have in the matrix ok.

So, let us look at the verbs part of that. So, they have taken 10 nearest neighbors for the verbs that are listed here. Let us take one, I am going to look at the understanding here. So, if you look at this, you have to comprehend, explain, understood, realize, grasp, norm, belief, recognize, misunderstand, understands and so on. I think this has done a good job here too. And then let us look at the adjectives part. What should I pick up? Let us look at the frightening part. So, scared, terrified, confused, frustrated, worried, or let us look at something else it is because, it found this very well. Let us, I am just looking for something which is not really a good one.

I am not able to find that. I think all adjectives have really good neighbors around here alright. So, with the set of nouns, with the set of verbs and adjectives COALS seem to perform better than HAL right.

(Refer Slide Time: 10:22)



COALS - SOME INSIGHTS

- The majority of the correlations are negative [4] ✓
- Word with negative correlations do not contribute well to finding similarity than the ones with positive correlation
- Closed-class words (147) convey syntactic information than semantic - could be removed from the correlation table ✓

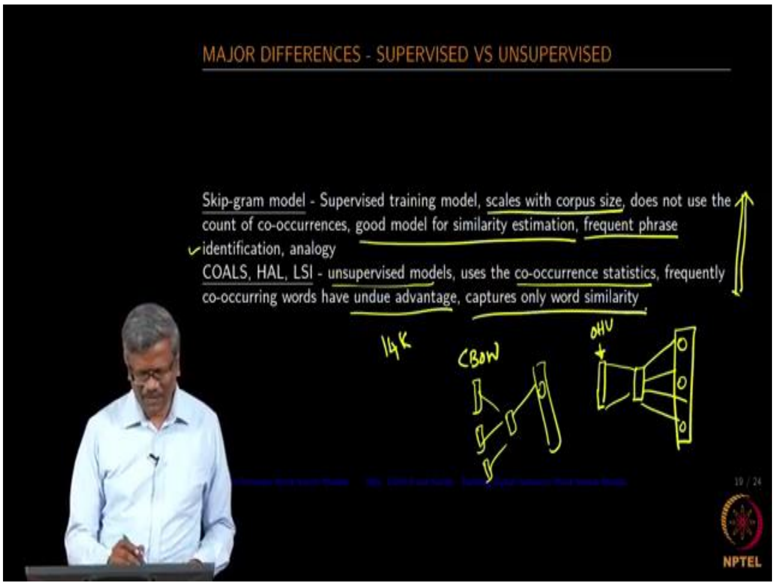
punctuation marks, she, he, where, after, the, 147

[4] Based on a Usenet corpus of size 1.2 billion words

So, in this case, the word vectors are either rows or columns; either one of them ok. One thing that they found was; the majority of the correlations are negative; that means, there are a lot of words that are not correlated at all in this space. There are a very small set of values that are available which are positive and those positive values represent the correlations for us. Words with negative correlation, do not contribute well to finding similarity than the once with positive correlation. The closed-class words of 147 convey syntactic information than semantic.

So, it could be removed from the table. Once are the punctuation marks she, he, where, after, the, and so on. So, there are they found close to 147 of them and they removed that. So, they are really not really contributing anything to the word vectors. So, in this case they have used the use net corpus of size 1.2 billion words ok. So, again by looking at the process of computing the matrix it is a little complex than what we saw with HAL. But, the outcome is really encouraging with respect to finding similar words right.

(Refer Slide Time: 12:10)



MAJOR DIFFERENCES - SUPERVISED VS UNSUPERVISED

Skip-gram model - Supervised training model, scales with corpus size, does not use the count of co-occurrences, good model for similarity estimation, frequent phrase identification, analogy

COALS, HAL, LSI - unsupervised models, uses the co-occurrence statistics, frequently co-occurring words have undue advantage, captures only word similarity

So, what are the major differences that you find at this point in time with respect to the supervised versus unsupervised one? So, so far we have seen these 3 models LSI is truly global because it uses the context of the document as well. Whereas, COALS and HAL they use the local context of 10 words. The only advantage of these two is that they actually utilize the count of the co-occurrences, ok. The supervised model scales well with corpus size. I think even COALS and HAL will also do the same thing.

The skip-gram model does not use the count of co-occurrence that is one major difference. So, is it possible to incorporate the count of co-occurrence in the skip-gram model? So, can we give input that as one parameter inside the skip-gram model? So, you remember that right especially in this skip-gram model we have the word vectors are fed and then, we are estimating the context words, right.

So, here we are inputting the one-hot vector of that word. So, is there a way that I can input the count of co-occurrence into that think about it; so, either in this model or in the CBOW model, where we fit the context and only identify the central word. So, can we input the count of co-occurrence in this? So, what will happen if you do that? The neural nets are supposed to take multiple features as well read as part of the input, think about this, ok.

The experiment is all about combining various inputs and trying them out unless otherwise, you know something is fundamentally wrong in terms of combining two sets of inputs ok. So, I am not sure whether there is anyone who had tried that kind of approach in the literature ok. Let us get into this part.

So, the good model for similarity estimation, because we are using a supervised model. Then frequent phrase identification also is part of that. I did not really cover that part but skip grams really would be useful in terms of looking at the frequent phrases for example, the phrases which are not linguistic phrases or it is like you know the New York Times type of phrases or the Indian Express, the Hindustan Times and so on.

So, those are the phrases that skip grams also are good at identifying. They are also good at this analogy part. I am not really convinced with the analogy part because you train the network with a set of analogy and then expect the machine to output that; obviously, we will do a good job there. So, we need to find out whether it can be done in an unsupervised fashion. So, that is where the challenges.

So, with respect to the COAL, HAL, or LSI these are all unsupervised models. They use the co-occurrence statistics, frequently occurring words have an undue advantage in the HAL, we try to reduce that using the COALS model by utilizing only 14 k values. It captures only word similarity. So, we are so, far we are not able to prove that also captures the analogy part of that.