

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

Lecture – 91
Global Vectors - GloVe

(Refer Slide Time: 00:15)

THE GLOVE MODEL

- ▶ The Global Vector⁵ (GloVe) models the word vectors with the computed statistics of the co-occurrence count
- ▶ The authors of this model introduce the idea that the co-occurrence ratio between two words in a context are strongly connected to the meaning

Let X represent the counts of co-occurrence matrix. Every element of X_{ij} represent the number of times the word j occurs in the context of word i .

Let $X_i = \sum_k X_{ik}$ be the number of times any word appears in the context of word i

Let $P_{ij} = P(j|i) = \frac{X_{ij}}{X_i}$ be the probability that word j appears in the context of word i

The skip-gram model captures the co-occurrences patterns one window at a time while the Glove captures it using the statistics of the co-occurrences or how often the patterns occur together

⁵ J Pennington, R Socher, C.D Manning, "GloVe: Global Vectors for Word Representation," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532-1543, October 25-29, 2014

20 / 34
NPTEL

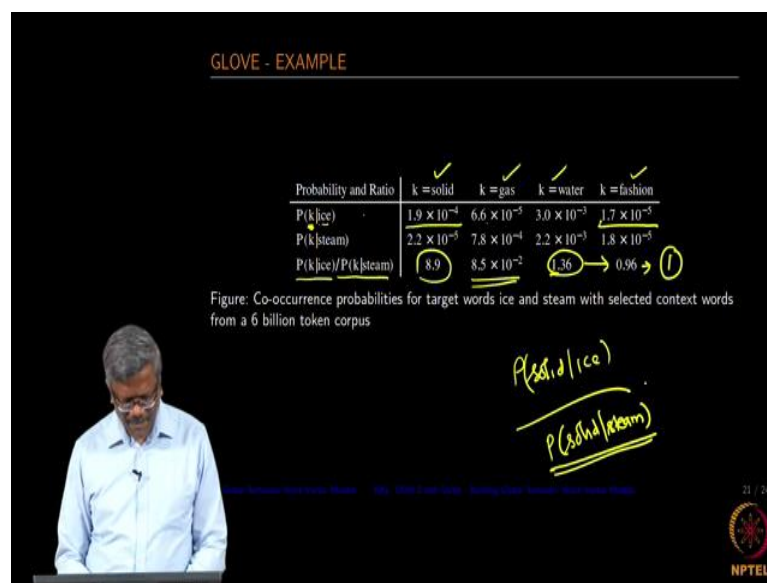
This paper is written by Pennington and others. This is available again for you to read. This small there is a very small section for analogy computation in this case especially for the Global Vector or the Glove Model. It is about one or two paragraphs I would like you to go and then read that. So, they use that as a downstream application to prove that the Glove model really solves the analogy problems, ok. So, now, let us get into the higher-level details of what the Glove model is. Again this paper is about this 8 pages long, ok. Except for the derivation of how they really construct the cost function rest are very easy to follow.

So, if you read that section in the third section where they talk about the Glove model you will understand how the cost function is constructed and then how they really use that cost function to find the word vectors. So, I am only going to give you a very high-level overview of this all right, ok. So, in this case they use to introduce the idea of co-occurrence ratio between the two words which we are not seen earlier. So, what we had seen earlier is only the co-occurrence count and then the correlation computation for two

words and then finally, we built the table right. In both in the Hal model where we utilize the count in the GloVe model we utilize the correlation right.

So, there is a definition that you may have to understand. So, this is the probability of the word j appear in the context of the word i is what we are finding. So, we introducing the concept of a probability in this case. In this case we are going to be utilizing these statistics of the co-occurrences or how often the patterns occur together. So, which skip-gram model failed to utilize.

(Refer Slide Time: 02:51)



So, to understand the particular slide here; you need to first go to the examples that they have given, ok. So, they are given to rather 4 classes let us say solid-gas water and fashion and then they have the probability computation forgiven the ice what is the value for this probability, ok when k is equal to solid or gas water or fashion, ok.

So, they use the same kind of the corpus in case, in this case I think they are utilizing a very big corpus and then try to find out the values for each of this, ok. So, to find the probability so, you need to find out the X_{ij} ; X_{ij} represents the number of times the word j occurs in the context of the word i . And then X_i is equal to the sum of X_{ik} , where X_i is the number of times any word appears in the context of the word i .

So, you take any word and then how many times any word appears in the context of that particular word is what you are computing for X_i here. And, then probability uses those

two values which are defined as the probability that the word j appears in the context sorry, appear in the context of word i , ok. So, this is what is being computed here right. Using the corpus we can compute these values.

So, what also they are suggesting is; if you want to really distinguish the meaning of two words this is not just enough you can also take the ratio of these probabilities that you have here, ok. So, let us first find out what this one is. So, as I mentioned earlier we have 4 classes and then the values are found using the corpus ok. The probability of solid given ice is this number, ok. And, then you replace that with gas and you get this number, you replace this with water you get this number and then you replace this with fashion which is not connected to this so; obviously, this is going to be a very small number right.

So, when it is very close right so, when k equal to solid you have a higher value, and given the ice you know it cannot be a gas right. So, this is going to have a small value and then this is also going to have a value higher than this in some cases, because of the occurrences and definitely this is not related to this so, you have a very small value. So, in the same fashion, you find it all ok. And, then you also try to find out the ratio or the probability. For example, if you look at this; since this is going to be very small we are going to have a very high value for that, right.

And, then again for the second one, it is going to have a very small value. Some of them which are not related to that could be having a value closer to 1. So, that means, it is possible for you to distinguish ice and steam easily in this case semantically you can separate them outright.

So, when compared to the raw probability the ratio is able to distinguish the relevant word; solid and gas from the irrelevant word water and fashion, ok. So, in this way you would be able to distinguish two different sets of words. So, that is what they are trying to prove by providing this example, all right. Since the ratio that we are talking about depends on 3 words right.

(Refer Slide Time: 08:09)

GLOVE

Since the ratio $\frac{P_{ik}}{P_{ij}}$ depends i, j, k , it can be modeled by $F(w_i, w_j, \tilde{w}_k)$. There could be several possible ways to encode the ratio. We would like to estimate the parameters of this model given the ratio.

Using the factoring approach similar to LSA, the new weighted least square regression model is proposed that minimizes the cost function

$$J(\theta) = \sum_{i,j=1}^{|V|} f(x_{ij}) (w_i^T w_j + b_i + b_j - \log x_{ij})^2 \quad (1)$$

where $|V|$ is the size of the vocabulary and

$$f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^\alpha & \text{if } x < x_{max} \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

The cutoff $x_{max} = 100$ and $\alpha = 0.75$ (4)

So, we have ice, steam, and a solid correct. So, it depends on 3 words, it can be modeled using a relationship going to be estimating a model which actually tries to get this ratio, ok. So, what we are computing is only this correct; so, we have the value given, but we do not know what that what created those values, let us say like this. If you know to see for example, in the case of 5 plus 5 you know it is 10, it is very easy to figure out, but if you are not given this so, there are so many different possibilities by which you can obtain 10. Is given in the formula

s

$$J(\theta) = \left(\sum_{i,j=1}^{|V|} F(x_{ij}) (w_i^T w_j + b_i + b_j - \log x_{ij})^2 \right)$$

So, what is the right model that gives you this 10? Based on various measurements that we have using these statistics, we should be able to estimate that particular model then that is what these authors try to figure out. So, what they did was they created and cost function that dependent on bias values and the word vector.

So, these are the 4 parameters that are going to be estimating and these values are available from the corpus, ok. And, then they also have a weighted function $f(x_i)$ and then they also define the weighted function as follows as given in this equation 3. And,

then they also have given the value as x_{max} equal to 100 and α equal to 0.75. So, there is no reason why only these values are picked, but they found that these two values really are giving an optimized vector set. So, that is why these two values are chosen, there is no theory behind these particular selections of values ok.

(Refer Slide Time: 10:42)

GLOVE

Since the ratio $\frac{P_{ik}}{P_{ij}}$ depends i, j, k , it can be modeled by $F(w_i, w_j, w_k)$. There could be several possible ways to encode the ratio. We would like to estimate the parameters of this model given the ratio.

Using the factoring approach similar to LSA, the new weighted least square regression model is proposed that minimizes the cost function

$$J(\theta) = \sum_{i,j=1}^{|V|} f(X_{ij})(w_i^T w_j + b_i + b_j - \log X_{ij})^2 \quad (1)$$

where $|V|$ is the size of the vocabulary and

$$\theta \Rightarrow w_i, w_j, b_i, b_j \quad (2)$$

$$f(x) = \begin{cases} \left(\frac{x}{x_{max}}\right)^\alpha & \text{if } x < x_{max} \\ 1, & \text{otherwise} \end{cases} \quad (3)$$

The cutoff $x_{max} = 100$ and $\alpha = 0.75$ (4)

So, in this case the authors are trying to minimize this function and θ so, these 4 parameters that you want to estimate using this particular model, ok. So, the algorithm tries to minimize this cost function, and then initially these values are randomly chosen as we progress based on the values that we obtained from the corpus, we keep improving these values. And then when it is closer to the minimum we stop the iteration or if it is here at the end we stop the iteration process, ok.

(Refer Slide Time: 11:44)

GLOVE - TRAINING

- ▶ Corpus size
 - ▶ 2010 Wikipedia dump - 1 billion tokens ✓
 - ▶ 2014 Wikipedia dump - 1.6 billion tokens ✓
 - ▶ 2014 Wikipedia dump + Gigaword5+ Common crawl of web - 42 billion tokens
- ▶ Input matrix size $X \in R^{V \times V}$
- ▶ Vocabulary - 400K frequent words →
- ▶ Initial learning rate - 0.05 → *Adagrad*
- ▶ Context words to the left = 10 ✓
- ▶ Context words to the right = 10 ✓
- ▶ Generates two words vectors \underline{W} and $\underline{\tilde{W}}$
- ▶ The final word vector = $\underline{W} + \underline{\tilde{W}}$

23 / 24

NPTTEL

And let us look at what they have chosen as their a corpus. They are tried with 3 different corpora; one is the 2010 Wikipedia dump which is about 1 billion tokens and then they have used the 2014 Wikipedia dump with 1.6 billion and then there is 1 called Gigaword5 that contains about 40 plus billion tokens. And, then they utilized a common crawl of the web and then used 42 billion tokens to train this model, ok.

And, then they have chosen close to 400K frequent words as vocabulary, I started with the initial value of 0.05 as the learning rate and then they used to Adagrad; as a mechanism to change the learning rate as they move along. And, then the context words to the left is considered and the context word to the right also is considered in this case they get 2-word vectors as in the case of Hal coal, as in the case of Hal and coals and then the final word vector is chosen to be the sum of those 2 vectors.


(Refer Slide Time: 13:24)

GLOVE - RESULTS

- ✓ $SVDL = \log(1 + X_{ij})$
- ✓ $SVD\sqrt{S} = \sqrt{X_{ij}}$
- HPCA: PMI version of LSA (PCA)
- ✓ \sqrt{LBL} , \sqrt{LBL} : log-bilinear model

Model	Dim.	Size	Sem.	Syn.	Tot.
\sqrt{LBL}	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
\sqrt{LBL}	300	1.5B	54.2	64.8	60.0
\sqrt{LBL}	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW [†]	300	6B	63.6	67.4	65.7
SG [†]	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
*GloVe	300	42B	81.9	69.3	75.0

24 / 24



And, then they have shown through their results that this performs better than any other model. So, they used SVD using a log, and then they also used a square root of the X_{ij} for SVD computation. And, then try to find out how good their model is with respect to various other models.

So, in this case there is one more which I have not discussed which is again on the latest which uses the PMI version of the LSA to find the word vectors. And, then there are two other models which are logged by linear models apart from what we have discussed earlier, they tried all those models and then try to compare Glove with this.

They started with 100 dimension vector and then the accuracy of their word vector is about 60 percent, ok. And, then they used a 300 if you look at this you know as in every paper you will see their model performing better than any other model. So, so far at this point in time Glove is supposed to be giving the best word vector and then the word vectors are also available for you to pick up and then use.