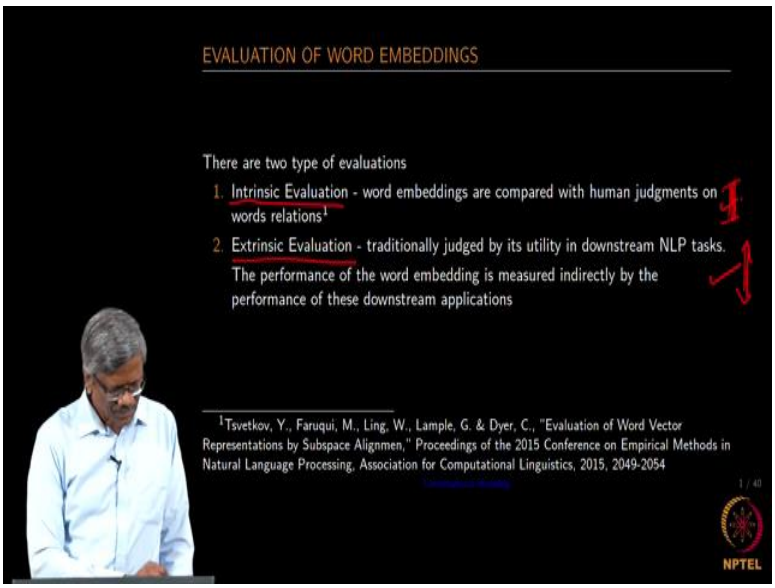
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture - 92
Evaluation of Word Vectors

(Refer Slide Time: 00:15)



We saw earlier two different approaches to creating word embeddings right, one is using the unsupervised model and the other one is using the supervised model. So, very early we talked about the CBOW and skip-gram models in order to create the word embeddings. And then later we spoke about two different approaches using coals and a half and then later using glow to create word embeddings, right.

Since, word embeddings are the most fundamental elements of any natural language downstream application. We need to make sure that we are creating a good word embedding vector or word vectors. So, in order for us to find out how good that word vector is we need to have some kind of a measure; so, the quality of the word vectors are extremely critical for the downstream application. If we do not have good quality vectors, we cannot expect a very good performance of those downstream application correctly.

So, we need to really have good quality word vectors and then we also need to figure out how to evaluate them. So, that you know we can measure the quality of those word

vectors, ok. So, how do we measure those word vectors? You know, there are two ways of doing it one is as I mentioned you know can input this as a vector for the downstream application and then measure the quality of the downstream applications. So, that is one way of doing it, the other way is to use an intrinsic model where we can compare the word vectors with human judgments.

So, this is the most important step right in order to really find the quality of the word vectors. So, one is to use the downstream applications that are called extrinsic evaluation. So, it is very important for us to really measure the quality of those word vectors using extrinsic evaluation.

So in this case how do we do it we input the word vectors as input and then measure the quality of the applications right, the downstream applications. So, if you keep feeding different sets of word vectors and then measure the quality of the downstream application and then find out which one gives you better results. So, in that way you should be able to measure the quality of the input word vectors.

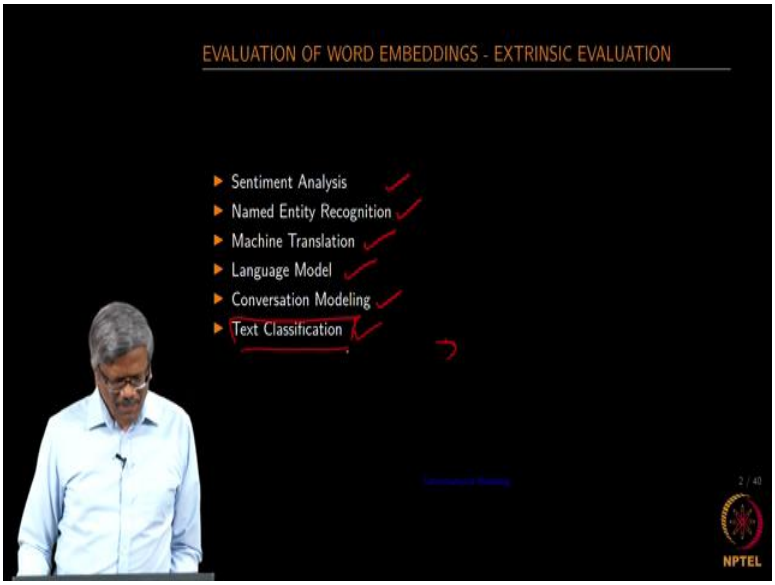
The other one is actually as i mentioned earlier you by using the intrinsic evaluation; so, in this case what we do is we compare the word vectors created by the applications with the word vectors that are judged by humans. So, what we do in this case is we give a set of words and then ask the human to verify whether the similar words are really applicable for a given word or not. So take one word and then provide various similar words to that and then ask the user to provide the score for each of those words, ok.

So, by that you are now going to have a set of vectors where humans have evaluated the entire vocabulary of words, right. And then once these applications like skip-gram or a CBOW model or any of those other unsupervised models create the word vectors find an automatic way to compare those word vectors to tell whether this particular set of words vectors are good or not, ok.

So, we have two different approaches that we are going to be looking at. So, we have already seen many applications in this case right, especially in the case of a sentiment analysis where we remember I used to take some vectors from the glove and then try to find out whether a given word belongs to the positive sentiment or negative sentiment from a given sentence. And then we used our language models in order to really find the in or predict the next word.

So, those are the downstream applications that we can think of, then measure the quality of that application and say that in the end the quality of those applications really depends on the quality of the word vectors that we have used as input right. So, this one which I have mentioned earlier we have not covered yet; so, we will talk about this in the next few slides.

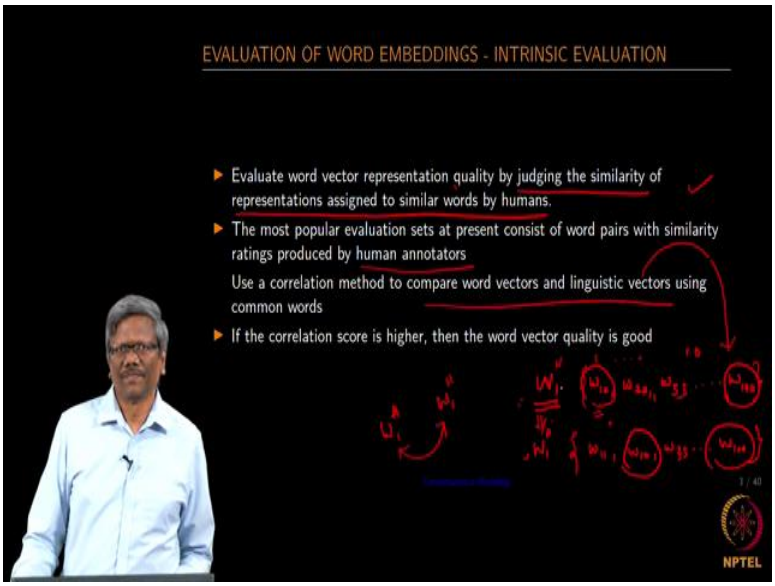
(Refer Slide Time: 05:17)



So, I think I spoke about this another application like sentiment analysis, named entity recognition also could be used as a downstream application to measure the quality of the word embeddings, machine translation, language model conversation modeling, and also text classification. I am sure you remember the classification atom that we have made using the BBC corpus, right.

So, we were not able to really get some good classification with respect to various classes that we have in that BBC corpus. So, can we use word vectors of different types and then find out whether the classification could be improved beyond a point is something that we can look at, ok. So, using this as well we can find the quality of the word vectors.

(Refer Slide Time: 06:19)



Ok, as I mentioned earlier so, we are going to be looking at the evaluation of the word embeddings using the intrinsic evaluation. So, in this case what we do is, we are going to be evaluating the word vector representation quality by judging the similarity of representation assigned to similar words by humans. So, what we do is let us say that we take a word 1 ok, and then based on the dictionary of synonyms are using wor let us assume that we are going to be getting.

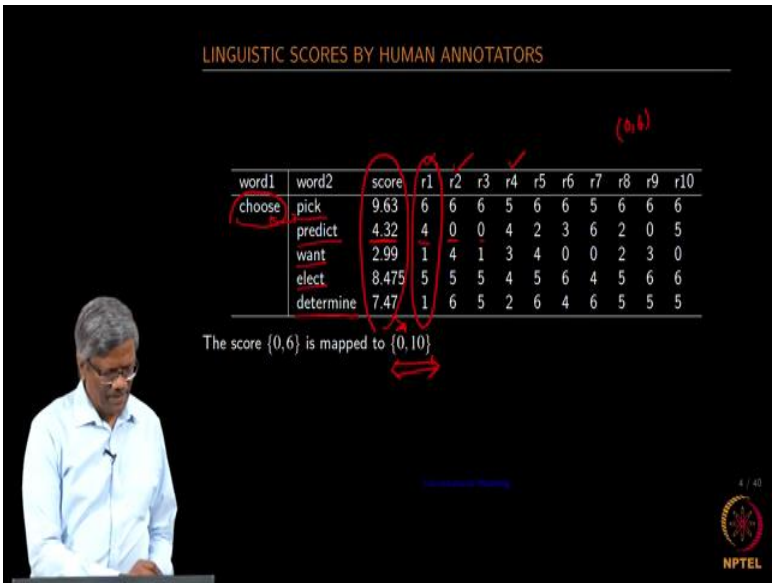
Let us assume that we have a similar word set for the given word w 1, ok and you provide this word set to the human and then ask them to rate each one of those. So, on the scale of 1 to 10 you can provide the values and then use those values to compare the word embeddings created by the automated systems. For example, this is the human evaluation; let us assume that for the same word vector coming through the automated systems. So, we get something similar to this, assuming that we have gotten something in this fashion.

So, the idea is to see how close this w 1 A to w 1 H. So, what kind of measures that we can use in order to compare these two is what we are going to be studying in the intrinsic evaluation of word embeddings. The most popular evaluation sets at present consist of words passed with similarity ratings produced by human annotators.

Use a correlation method to compare word vectors and linguistic vectors, the linguistic vectors are the ones that are created by humans using common words. So, you pick up a

common word as I mentioned here and then try to compare it or use a correlation method. So, that you get some score for w 1 A. So, if the w 1 A is pretty close to w 1 H then we can say that we have obtained a good word vector from the given model.

(Refer Slide Time: 09:08)

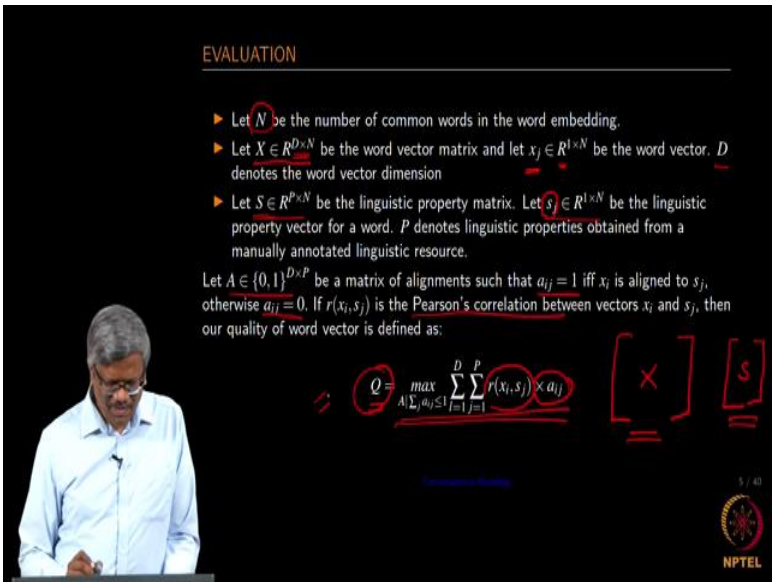


So, this is one of the examples that I am showing here in this slide for example, here choose is the word that is chosen and then there are several words given as a word 2. And then it the authors have asked the human annotators to give a score for each of those so, how to choose and pick are similar so, what is the rating that you would provide.

So, one person has given 6, the second one has given this as 6 and so on except for this all ratings are 6. So, in this case the ratings are from 0 to 6 and then for choose and predict ok. So, this is the mean score we have 4 0 0, and so on, like this you know you provide all the word combinations and ask the human annotators to provide the score for each of those and every annotator. So, you know without referring to anyone else had scored this. So, that should be the idea right.

And then, what they do is at the end when these scores are computed it is mapped to 0 to 10 and this is the scale most of the evaluation mechanism would use ok, for pair of words you will have some score in between 0 and 10, ok

(Refer Slide Time: 10:45)



So how do we do that so, it is now having said this in several words let us try to define some notations, ok. Lets N is the number of common words so, this is a very important right. So, we need to have a set of common words that are available in both automated as well as in the linguistic set. And the next belongs to the space of D by N right. It is a real space whose dimension is D by N, where D is the total number of denotes the word vector dimension I am sorry. So, D denotes the word vector dimension, and then for each word in that you have you are going to be having that in the same real space and whose dimension is one by one, ok; so, that many the columns you will have for this. Is explained with the equation below

$$Q = \max_{A \mid E_j a_{ij} \mid \leq 1} \sum_{i=1}^{D} \sum_{j=1}^{p} r(x_i, s_j) \; X \; a_{ij}$$

And then again we have the linguistic set defined like this and then each word we have this one is the linguistic property vector and then P denotes the linguistic properties obtained from a manually annotated linguistic resource. So that should be the size of that ok. And then what next we should do, right. So, we have the word embeddings coming from the system from CBOW or skip-gram models and then we have the linguistic scores obtained by the human annotators. So, the idea is now to compare that right so, when you get the scores arranged in a certain fashion.

So, we need to start aligning them ok; that means, we want to find out for the given word whether there are any words found by the CBOW or skip-gram models, ok. And if that particular aligned word should also be found in the linguistic vectors, if it is not found then the alignment is going to be a 0 otherwise it is going to be 1; that means if you have those vectors as I mentioned earlier, right. So, there should be at least 1 alignment like this.

So, we have some alignments so; that means, it is important to understand that the similar words obtained by the word vectors need not be the same as what the linguists have provided, ok. So, if you have this right; so, you have two different sets of vectors x and then you have one more from the linguists side that is equal that is S. And then when you do this correlation operation across each word that you have in X as well as in this you get a matrix of this type. You try to find the value Q using the correlation and the alignment that you have. So, this is the score that you want to achieve for the evaluation, ok. So, this is one of the methods of evaluating word vectors.

So, you have a set of vectors coming from the linguist which are manually tagged and then we have the word vectors coming from the skip-gram model or from the CBOW model or from hal or from coals or from the glove and then use a correlation method for the common words that are found in both linguists matrix as well as the matrix obtained from these methods. I have used a research paper to quote the idea, there are several other mechanisms to find the or several other mechanisms to evaluate the quality of the word vectors as well, ok. With this, I close the evaluation of the word vectors and also the lecture series.

So, I like to thank you guys for being part of these series throughout the 12 weeks and I also would like to sincerely thank the team the NPTEL team, who is behind all these video recordings, assignments, releasing the videos on time and so on ok. I also like you to take your time to send a thank-you note to these wonderful folks.