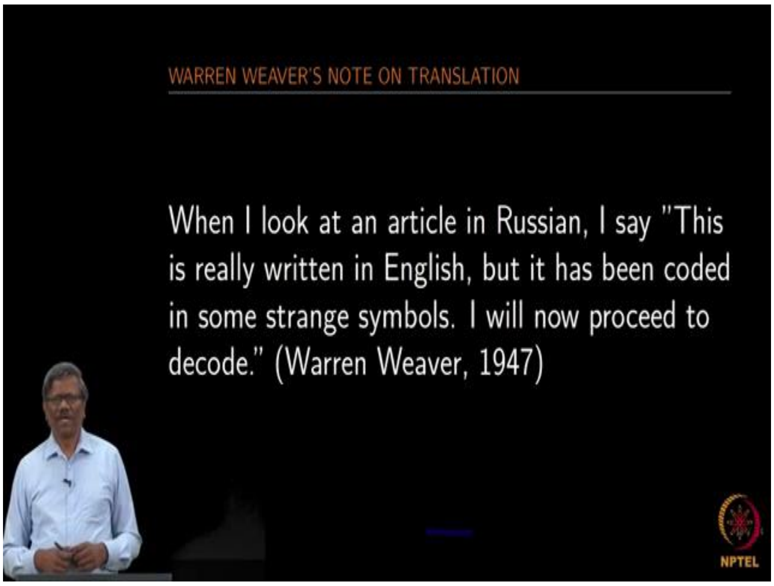
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture – 06
Machine Translation

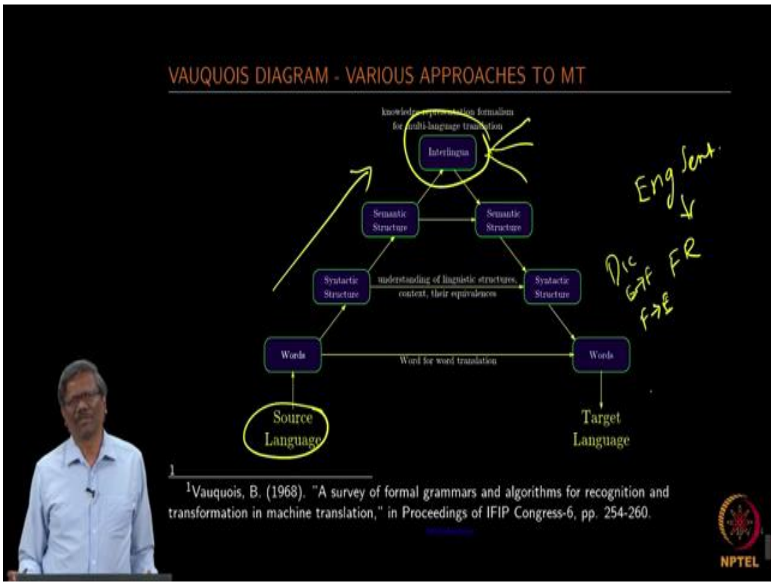
(Refer Slide Time: 00:15)



So, next, once you have understood a sentence, you are able to construct the sentence in the form of a fixed-sized vector. Can I use it to translate? Ok, that is the next question right. again the translation is not a very new problem, it is been therefore they it is been there for more than 60 plus years ok. this is what really started the machine translation process.

So, this is by one Warren Weaver a researcher; in 1947 he said, when I look at an article in Russian, I say this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode ok. this is what really started the machine translation in a big way. after this you know people started looking at various statistical forms of freely converting the language into from one, really translating the language from one to the other. And then they started evaluating methods, designing new models and so on.

(Refer Slide Time: 01:39)



This is something that I thought I would bring in for you to understand this, you know when you start the translation process, you look at, for example, you have a dictionary, let us say you have the English sentence at source, and then you want to translate it to French. You have a dictionary English to French dictionary, and the French to English dictionary right.

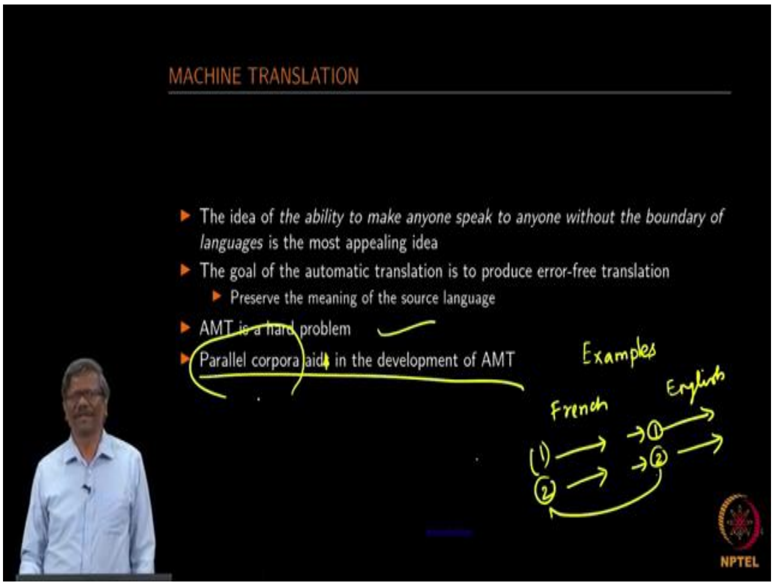
So, if the English sentence is given what you do you just look at that word ok, go to the English to French and dictionary and find the translation of that, and then write it. word by word, you want to translate. I do not think it will give you good translations. The especially if you translate from let us say English to any of the Indian languages, it is going to be really hard if you do a word by word translation. This sentence would not be very meaningful.

Then what they did, they started creating a syntax structure. For example, English has a structure in terms of forming the sentence right. And then there is an equivalent structure in the other language to which you want to translate into copy from the syntax, from the syntactic structure of the source into a similar structure on the target language and then do some translation, so this is little better, but not quite.

And then you need to have this semantic understanding of that you create the semantic structure. these are all the hypothetical think that they started looking at and then they found that there could be some mechanism, where I can keep the source language into interlingua or the knowledge representation of the source language, which I can use it to translate into various other languages. this is a certain theoretical approach people started looking at and they were not very successful at that ok, because it requires a lot of making capability as well.

So, I am sure you know that in those years, the computer power that they had is a lot smaller than what we have on our wristwatch right now, correct. the digital one that we are talking about right now has a lot more power than the computers that they had in the early 50s and 60s.

(Refer Slide Time: 04:40)



So, they were not able to move beyond a certain point. we started looking at theoretical approaches that are appealing to people ok. this translation is a very very interesting idea that you have right; assuming that there are folks who have come from a different country and do not understand English or your mother tongue ok.

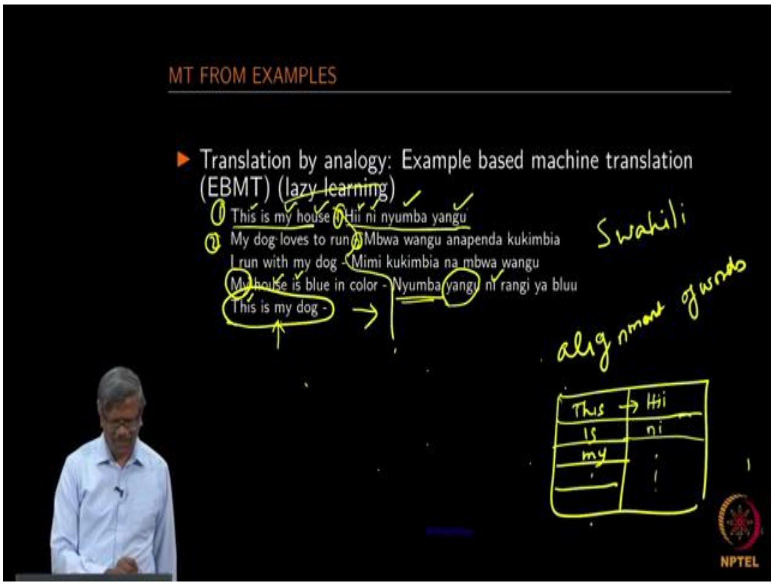
And you do not understand their language and if you want to communicate with them in a certain fashion, you know you can use some gestures to communicate, but probably may be good enough for certain small tasks and so on, but not good enough to communicate certain great ideas.

So, is there a way that we can create mechanisms and when you speak in your mother tongue, it automatically translates that into the language that the other person understands and he listens to that voice in his own mother tongue. He understands and speaks in his own mother tongue and you get to hear that in your mother tongue, when the translation finally happened right. in this way you can communicate there could be some translator sitting in front of you, who is doing that job and if you do it in real-time that will be really great that is actually the goal of the machine translation that we are looking at ok.

It is a very hard problem we are trying to attempt this, so for that, we require some parallel corpora. Parallel corpora are plural, so we require parallel corpora which is nothing but sample or example sentences, so you have French, so you have 1 sentence here and the equivalent of that is available here as 1 ok.

And then you have translated versions available and they are available 1 is corresponding to 1 in, the sentence 1 in French is corresponding to sentence 1 in English; 2 is corresponding to the French sentence 2 here and so on and so forthright. we need to have a huge corpus of the translated version, so that is what we call as parallel corpora. it should be available for the machine to learn as we had seen earlier right. we need parallel corpora in order to really understand the translation process.

(Refer Slide Time: 07:29)

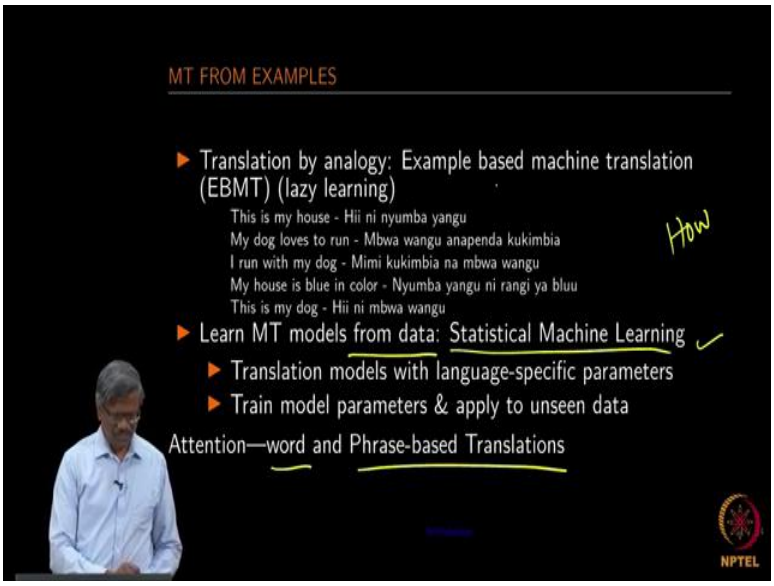


So, we can do this by analogy or you can also use or lazy learning mechanism to understand a language. For example, this language that I have here on the right side, a Swahili and this is English. by looking at these parallel sentences ok, so I used one of the translations that are available online and then I got this ok. this is my house and then this is the equivalent of this sentence in Swahili ok, so we have 1 here, this is 1, 2 and 2. By looking at the words ok, so would you be able to translate this is my dog into Swahili, I think to some extent it should be possible right. this is available so wherever you have this you can say ok.

And then there is ease here and then knee could be corresponding to that and then my corresponding to this and then the house could be this, because if you look at this and here and then or it could be the other way because we do not know; but roughly you can translate this particular sentence into Swahili if you know or if you have parallel corpora in this fashion. this is so very rudimentary way of doing it, but fundamentally that is what is being used and the machines for the translation it.

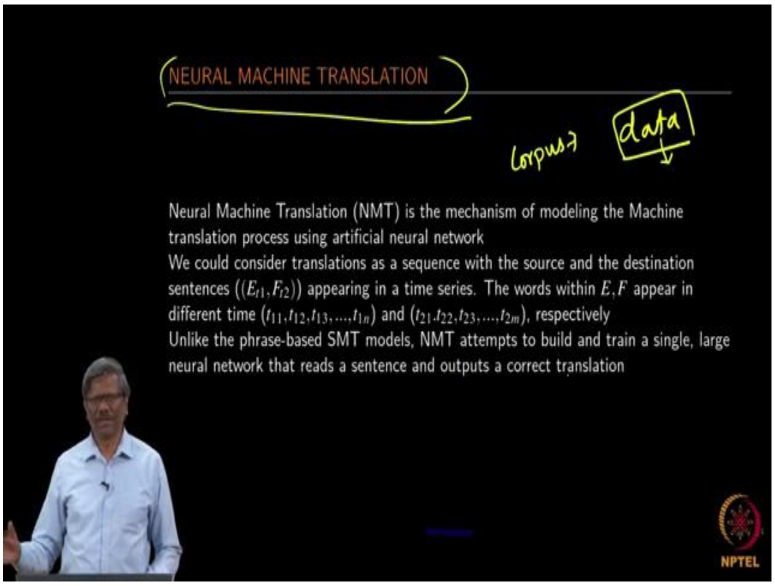
Tries to look at some parallel or alignments what we call as ok, so we look at the alignment of words and then try to see if we can form a new sentence given an English sentence. using the alignment as a table, so I can just create a table like this and then start figuring out you know how many times this occurred in English. And then how many times that particular word occurred you know that will give you some idea of you know this is corresponding to let us say Hii in this case and then is ni and so on. And then using this table I can translate this sentence into Swahili ok.

(Refer Slide Time: 11:05)



So, we want to learn those models from the data that is available in the parallel corpora. we have two approaches that we will talk about one is the statistical a machine learning approach ok, where we use certain language-specific parameters in terms of alignments. And then we will also try to figure out there could be phrases right not just one word, there could be phrases that could also align from one language to the other. we use two models to do machine translation using statistical models. we use the word and phrase-based translations. And then this is one way of how can I do the translation.

(Refer Slide Time: 11:57)

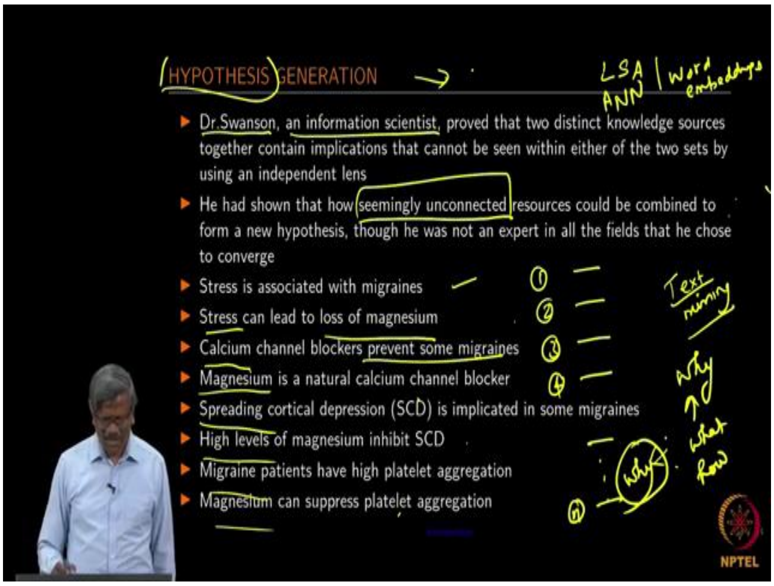


And then since we had seen earlier that it is possible for the sentence for the neural net to really code a sentence or encode a sentence into a fixed set sized vector. can that fixed-sized vector could be used for translation is what machine learning all about in the translation space. there also we look at word-based models and also we looked at alignment-based models or attention-based models. we will talk about that in detail during that lecture ok.

So, look at this you know the way we are progressing, you might have now noticed that we are not going to be looking at the natural language processing from the syntactic structure that is available, you are going to be purely looking at the natural language processing based application from the data alone. for us corpus is the key, so can I learn anything from the corpus. everything that we have seen so far right, right from the beginning in terms of counting the number of words in terms of fearing out how the word embedding is obtained and then how can I encode a sentence, so all of them are coming from the corpus site.

So, we are not going to be using anything from the syntax side of the language ok, so that you should be aware of in this course we will only be dealing with the natural language processing, and then we will only be taking the information from the data and nothing more ok. You are not going to be considering any grammar syntax structure or any of those throughout the course ok. again the purpose of this is to translate from one language to the other using the neural network.

(Refer Slide Time: 14:00)



Last, but not least is the hypothesis generation, this is again it is a very important thing that we have not to scene yet I know good progress in the natural language processing. this is something that started the text mining activity ok. I will just talk about what this person did Dr. Swanson was an information scientist, who used to be sitting in the library helping researchers ok. his job was to really look at the keywords given by the scientist, and then bring in all the relevant documents and then provide arraying them in a certain fashion and then give it to the researchers that is exactly what he was doing in the library as an information scientist, but he was bit curious then what an information scientist.

So, he was a lot more curious than any other person as a human right, so that sees nature. You start looking at why are these people looking at various documents and then let me see something into it. Even though he is an information scientist, he started looking at various aspects of what he is retrieving and then providing me with the research community ok. While doing his activity, he actually had shown that it is possible to connect seemingly unconnected resources and form a few hypotheses. he was actually reading all the documents or the research papers that he is retrieving for the scientist and then start looking at seemingly unconnected concepts ok, as 1, 2, 3, 4 and so on ok, so that is the curiosity right.

So, he was very curious to find out if these unseemingly connected documents, have any connection at all so that is what he was looking at. actually he was looking at a few concepts that come into his repository. he started looking at see for example, some of the scientists were looking at stress-related to migraine. he retrieved documents that are related to stress in migraine and gave them those documents. He also started looking at rather there are certain scientists, who were looking at stress and loss of magnesium as another research idea.

And then has other scientists started were looking at calcium channel it is another scientist started looking at calcium channel blockers prevent some migraine ok. this seemed totally unconnected right when you look at this, so likewise there are several ideas research ideas that he is helping the research scientist’s way in terms of retrieving the document related to them.

So, when he is started looking at this he found that each one is some way connected to the other. And then finally, he came up with a hypothesis saying that magnesium deficiency could be one of the reasons for migraines. it could be one of the reasons there you know there could be several possible reasons for that that is one of the hypotheses that he had come up with. And then he actually gave presented that too few scientists and they try to really look at the details of what he has presented and there was a research paper that claims that what he ever whatever he had found is possibly true, so that means he is able to really connect certain things which are seemingly unconnected by just looking at the contents of the research papers right.

So, can this be replicated using the machines you know these as we progress right this kind of so as we progress in the natural language processing we have to start moving the research towards those directions? this is something which I want you to look at you know, there I do not think there is any model that are available which can generate hypotheses based on the collection of the document that it fights ok, and it requires really the domain knowledge in order to really connect those documents right. this is one of the very high-level intelligent activity that we want the machine to perform, especially in the natural language processing.

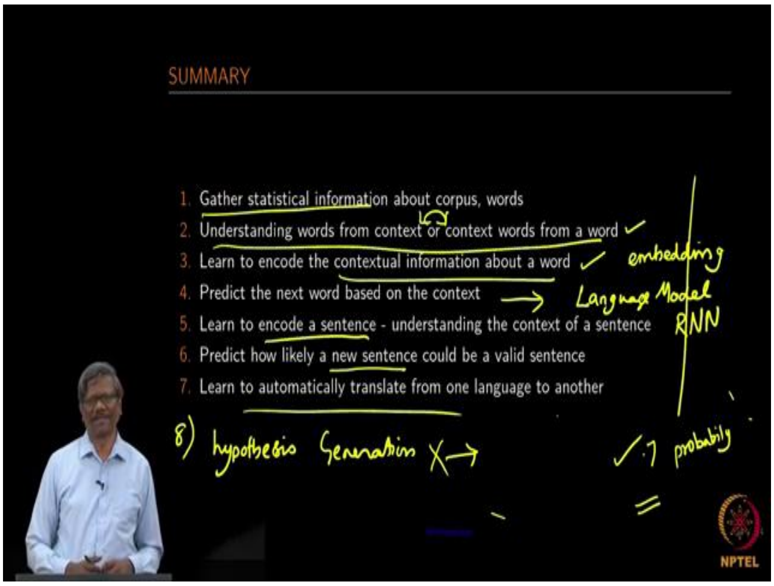
So, so have you seen the progression now of how we have moved from just counting the words in terms of the frequency, in terms of inverse document frequency, and then later trying to find out what or the patterns in then, and then try to find out using the patterns what could be the next word for the probability of the next word following certain words that the prediction we try to bring in. And then later we try to really understand the context in which the words are presented in the corpus. And then later try to use that to figure out whether a sentence could be formed by combining multiple words or string multiple words together to form a sentence.

And then we later we try to use various models to generate those you know one is using the LDA I am sorry, LSA to get the word embeddings or using the neural networks to get the word embeddings ok. And then later try to form a vector based on these sentences, and then try to do the translation and finally, we are trying to figure out whether a hypothesis can be generated by looking at a collection of research documents and so on right.

So, I keep repeating this, because this is the core of what we are going to be looking at in the next 12 weeks. also remember that whether we are really answering, why using what and how ok. So then why should we be at your center, so this is central to do it why are we doing this ok. the reason why we are doing it at a very high level is to give the machine the ability to do what we do in an intelligent fashion ok.

So, by doing this slowly and steadily we have moved from just from the counter to learn certain things about the word, and then slowly of even when he comes into the translation and hypothesis generation, we are moving the learning into the deep learning aspects and then we start to provide deep learning models and so on and so forth. And then the deep learning models in turn really take us to the intelligent task that we perform as humans,.

(Refer Slide Time: 22:41)



In summary if you look at this, I think I have mentioned all of this. We initially keep repeating this one more time, we gather statistical information about the corpus words, understanding the words from the context or context words from a word, so both ways ok. understand the words from the context or given the word get me all the context words ok, so that is again would be part of our lecture series. Learn to encode contextual information ok, so the contextual information about a word is encoded in a certain fashion that is what we call as the embedding right.

And then predicting the next word, so that is where we bring in the language model. we try to build the language model which will try to predict the next word based on the context. And then learn to encode a sentence understanding a sentence, so he said we require sequence learning and then we use a model called RNN to encode a sentence. And then we use the same model to predict how likely a new sentence could be a valid sentence.

Suppose he inputs a new sentence and the machine should be able to say that ok, it is 0.7 as the probability that means, its this sentence is highly probable and it is well-formed as based on what it had seen earlier in the corpus, so that sentence could be passed as the sentence that is legally valid in terms of syntax. And then learn to translate from one language to the other. And then the last one is so I would not be talking about this because we do not have material later to that as of today ok. this could be a future topic, maybe in the next 3 years or 4 years I may be talking about this.