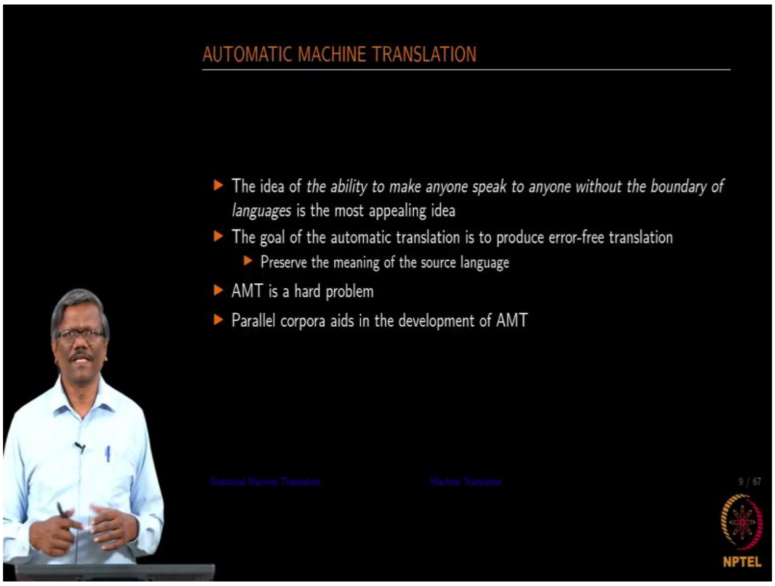
Applied Natural Language Processing

Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture - 64
What is SMT?

(Refer Slide Time: 00:15)



AUTOMATIC MACHINE TRANSLATION

- The idea of *the ability to make anyone speak to anyone without the boundary of languages* is the most appealing idea
- The goal of the automatic translation is to produce error-free translation
  - Preserve the meaning of the source language
- AMT is a hard problem
- Parallel corpora aids in the development of AMT

9 / 67

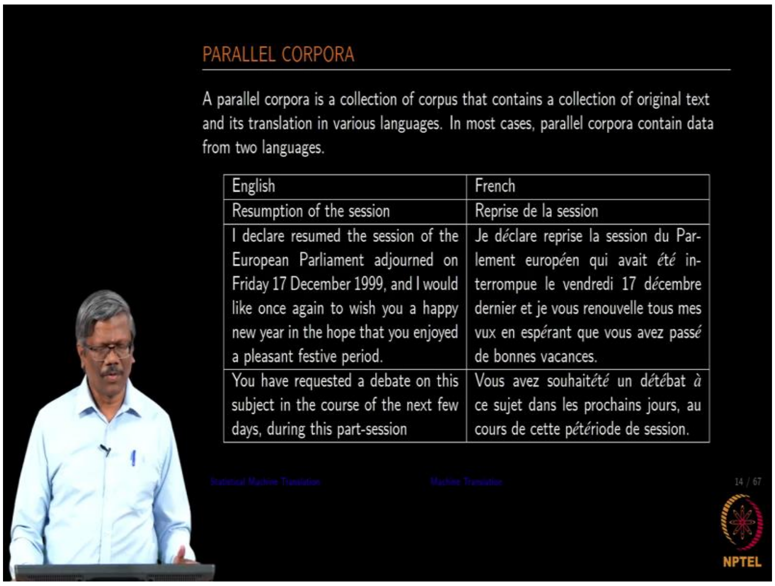
What really drives us now to get into the automatic machine translation or we call it as AMT. It is a very ideal thinking right. So, I speak in one language and then a person in Africa who only understands Swahili understand what I am speaking for example, he does not know my mother tongue and I do not know his mother tongue but I speak in my mother tongue and then when I pick up the phone and then call him for example, he is one of the best teachers for a certain subject and I want to hear from him about certain ideas that I have generated.

So, I call him in my mother tongue and then he understands what I am saying in his own language for their some let us assume that there is the black box that converts that and he speaks in his own mother tongue Swahili and then when I receive I receive the translated version in Tamil. So, I am able to really converse with that person in real-time.

So, it is a very ideal thinking; that means you have to break the complete boundary of language in the world if you are able to do this. So, this is the ideal situation and this is what we want to get to. I am not sure how long it is going to take us to get there but that is where people are working on and research is really getting into that direction alright. So, the most important aspect of any translation is to have error-free translations. So, when I translate I just want to really create an error-free translation without any semantic and syntactic errors associated with that; that means, we are preserving the meaning of the source language without adding any noise to that. So, now, you know that it is not a very easy problem. It is a very hard problem to solve ok.

So, how do you solve where do you start that is where a lot of parallel corpora there are available would help us. So that is a plural ok. So, if you look at the European Parliament they have translations from one language to the other whatever as happen in the parliament been encoded. For example, it is available from English to French, French to German, German to Spanish and then Spanish to whatever other languages that they speak there ok. So, this is something that we really need. So, what does this contain?

(Refer Slide Time: 03:21)



PARALLEL CORPORA

A parallel corpora is a collection of corpus that contains a collection of original text and its translation in various languages. In most cases, parallel corpora contain data from two languages.

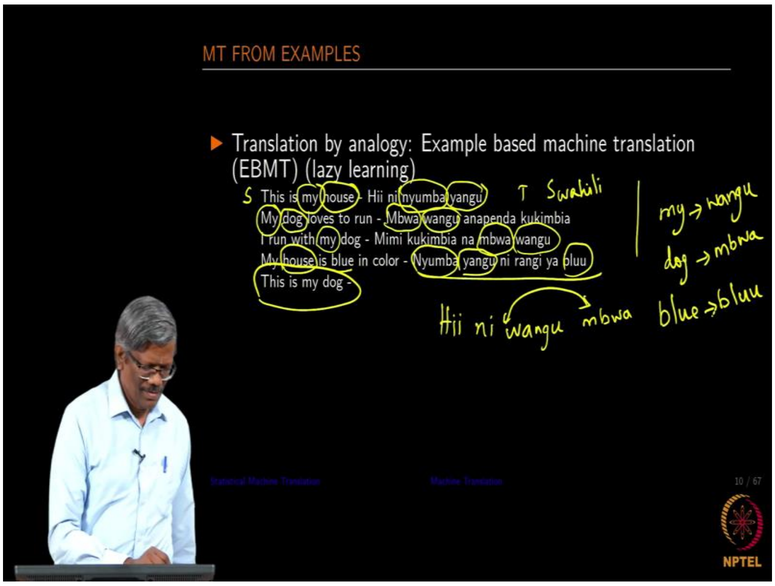
English	French
Resumption of the session	Reprise de la session
I declare resumed the session of the European Parliament adjourned on Friday 17 December 1999, and I would like once again to wish you a happy new year in the hope that you enjoyed a pleasant festive period.	Je déclare reprise la session du Parlement européen qui avait été interrompue le vendredi 17 décembre dernier et je vous renouvelle tous mes vux en espérant que vous avez passé de bonnes vacances.
You have requested a debate on this subject in the course of the next few days, during this part-session	Vous avez souhaitété un détébat à ce sujet dans les prochains jours, au cours de cette pétériode de session.

Statistical Machine Translation Machine Translation 14 / 67

Let me take you to that first. Parallel corpora are the collection of the corpus that contains a collection of the original text and it is a translation in various languages ok. Suppose if the language of communication in the European Parliament is English then they have the translated version of that in all languages. It is translated by humans. They have not used any automatic translation probably they would have used as the initial stage but the end all are corrected and verified in the manual fashion.

So, we assume that this particular corpus will give us the exact translation of one sentence to the other or from the source language to the target language. So, what does it contain; if you go on and look at their corpus it is all listed in pairs though. You can take English to French or French to English or French to German and then when you look at that file you will have 2 sets of a text file; one containing English other one containing French in this case and then the first sentence of the English text would correspond to the first sentence of the French. So, they are 2 different files. So, you have to really map them to create your own parallel corpora if you want to run it through your applications. So, they are 2 separate files the European Parliament data ok. Let us go back.

(Refer Slide Time: 05:09)



MT FROM EXAMPLES

- Translation by analogy: Example based machine translation (EBMT) (lazy learning)
  - This is my house - Hii ni nyumba yangu
  - My dog loves to run - Mbwa wangu anapenda kukimbia
  - I run with my dog - Mimi kukimbia na mbwa wangu
  - My house is blue in color - Nyumba yangu ni rangi ya bluu
  - This is my dog -

Statistical Machine Translation   Machine Translation   10 / 67

So, let me show you what can be done right. So, so far we have not looked at the language from the syntax perspective. We have been only looking at from the data perspective and we want to continue to do the same we do not want to get into the aspect of the grammar and things like that. So, we want to see whether the word embedding can be created using the data that is found we want to find out whether languages whether a sentence can be translated from one language to the other using the data that is available in the parallel corpora and so on.

So, how do we really start you know it is very interesting thanks to the efforts of the researchers that we have right now we are able to have an application that translates one from one language to the other. For example, I used a Google translation for this and we will learn how to really translate from English to the other language which we really do not know ok.

I have taken English as the source and then the target is Swahili. I do not know how to pronounce these words I do not know this language but I just want to see by looking at the words is it possible for me to understand what each word means in that language ok. Let us first look at one sentence. This is my house. So, pardon my ignorance in this language. So, I may be wrong in terms of pronouncing this; this is his ni nyumba yangu is the translation ok. So, now, from the first sentence we do not understand what corresponds to what; right.

So, there is no one to one there is someone to one correspondence, but we do not know whether the correspondence is correct ah. My dog loves to run mbwa wangu an agenda kukimbia ok. So, there is only 1 my here. There are 2 words; they are equal. So, we still do not find but there are certain things that sound similar.

So, we can assume that I could be corresponding to either of this yangu or wangu ok. I run with my dog Mimi kukiambia an MBWA wangu. So, now, there is third my around here. So, there is a range here. So, we can say for sure with a certain probability that my could correspond to wangu right. So, now, we have a second-word dog. We have this mbwa there is one more here. So, dogs may correspond to this right. So, you look at the third sentence my house is blue in color ok.

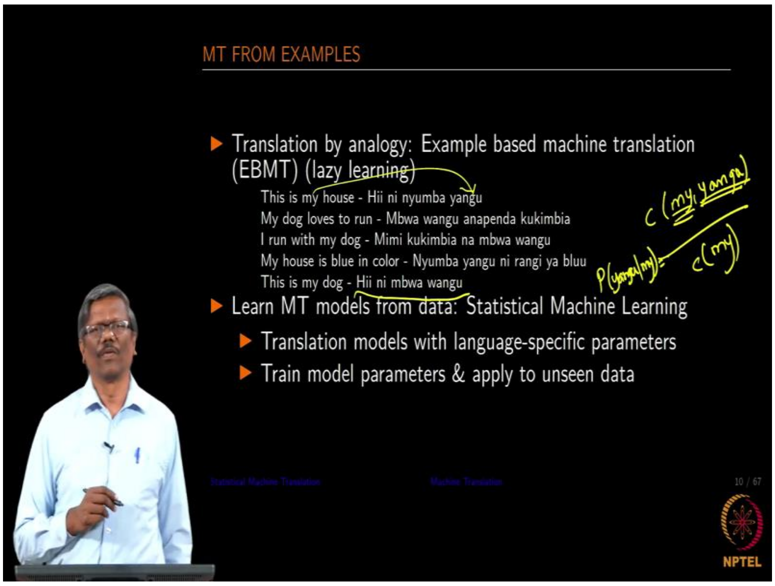
So, there is a translation available and then this blue you know by looking at this we can say that this could be ok. So, now, we know my and then here we found the house in the first sentence and then in the last sentence we also have this. So, this could be our house right.

So, it is possible for us to write. So, if you have the languages side by side and then we probably would be able to do the translation by analogy. It is could be you know we probably would be able to translate word by word but we may not be able to formulate a sentence based on this maybe if you have seen thousands of such translation we might be able to do the translation from English to Swahili tool right.

So, this is the idea right. So, now, we are able to really count how many times certain words are associated with the source to the target. So, once you are able to really find out that this particular word has been associated with this several times then we can for sure say in that the word in the source language is associated with this word in the target with certain probability ok.

This is the idea. This is how we are going to be looking at the corpora and then trying to find out the translation aspect of it and then try to formulate various algorithms to do this job alright. Actually, I just wanted to ask you this question; this is my dog ok. So, how do you translate this ok? So, we have seen that there is some h I I and then is maybe n I then my we saw either it is yangu or wangu say since it is associated with the dog here my there is a wangu there. So, maybe I can and then m b w it could be here or there ok.

(Refer Slide Time: 11:35)



MT FROM EXAMPLES

- Translation by analogy: Example based machine translation (EBMT) (lazy learning)
  - This is my house - Hii ni nyumba yangu
  - My dog loves to run - Mbwa wangu anapenda kukimbia
  - I run with my dog - Mimi kukimbia na mbwa wangu
  - My house is blue in color - Nyumba yangu ni rangi ya bluu
  - This is my dog - Hii ni mbwa wangu
- Learn MT models from data: Statistical Machine Learning
  - Translation models with language-specific parameters
  - Train model parameters & apply to unseen data

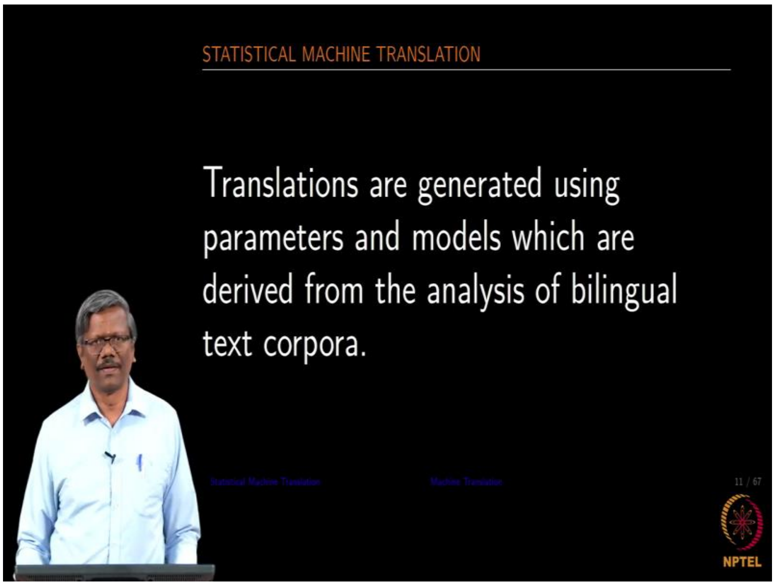
$$P(\text{yanga}|\text{my}) = \frac{c(\text{my}, \text{yanga})}{c(\text{my})}$$

So, if you use this Google translation it tells me that this is it. So, somewhat we are really able to translate from the source language to the Swahili depending on what data that we had seen earlier ok.

So, this is what we call statistical machine learning. So, we are able to really count how many times certain words have occurred and then how many times those words had from source to the target had been aligned and then create a count of the probability of that particular word for example, if I am associated with the young. So, we will have the probability estimation like or let us say this count let me put it straight how many ever times we had seen mine.

So, this will give you the probability of the translation for my given translation yangu given my ok. So, we can write it in this fashion ok. This is the model that we are going to be following in this session for machine translation ok. So, for us to train we need to find the model parameters and then later apply to the unseen data like what we had done here right.

(Refer Slide Time: 13:25)



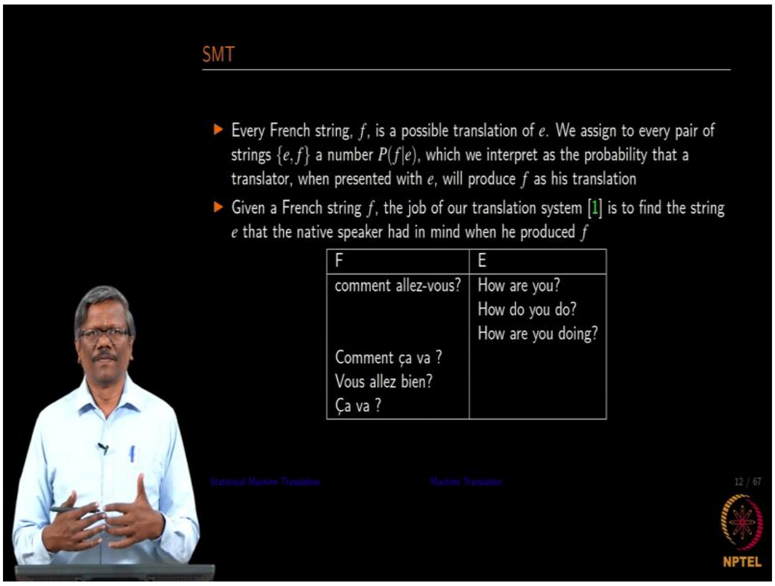
STATISTICAL MACHINE TRANSLATION

Translations are generated using parameters and models which are derived from the analysis of bilingual text corpora.

Statistical Machine Translation Machine Translation 11 / 67

So, this is the definition that I will be using statistical machine translation is defined as follows. Translations are generated using parameters and models which are derived from the analysis of bilingual text corpora. So, by looking at the parallel corpora we going to be generating the parameters and then using the parameters we will develop a model and then using the model later when we input a sentence that is not been seen earlier it would give us a translated version and then we can use some kind of a gold standard to find out whether the translated version is really a legal sentence or not a legal sentence alright.

(Refer Slide Time: 14:16)



SMT

- Every French string, f, is a possible translation of e. We assign to every pair of strings $\{e, f\}$ a number $P(f|e)$, which we interpret as the probability that a translator, when presented with e, will produce f as his translation
- Given a French string f, the job of our translation system [1] is to find the string e that the native speaker had in mind when he produced f

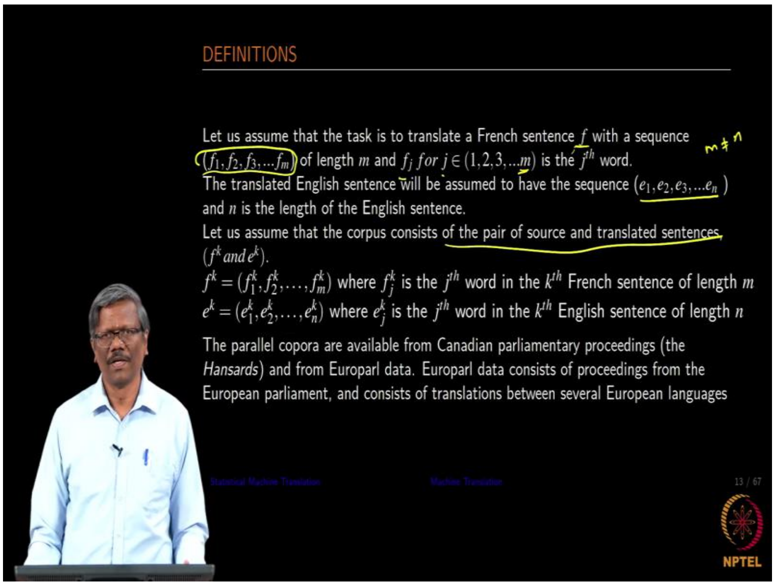
F	E
comment allez-vous? Comment ça va ? Vous allez bien? Ça va ?	How are you? How do you do? How are you doing?

Statistical Machine Translation Machine Translation 12 / 67

So there is another thing that I want to talk about here. For a given sentence you know there could be various ways of saying the same in a target language. For example, it is a very archaic way of saying in French comment Allez-Vous. You can in English you can translate that into how are you, how do you do, how are you doing. So, the same way you can comment ca VA or Vous Allez Bien, cava. I do not know whether that is the right pronunciation ok.

So, all it can be translated into multiple sentences that mean the same. So, it is possible for you to translate the source language into multiple sentences into the target language also. So, what we want to find out is we want to find out which one is the which one is likely to be translated in the right way by looking at the probability of each one of those sentences or which over one has the highest probability then that we will choose as the translated sentence for the source that we had provided alright.

(Refer Slide Time: 15:46)



DEFINITIONS

Let us assume that the task is to translate a French sentence f with a sequence $(f_1, f_2, f_3, \ldots f_m)$ of length m and f_j for $j \in (1,2,3,\ldots m)$ is the j^{th} word.

$m \neq n$

The translated English sentence will be assumed to have the sequence $(e_1, e_2, e_3, \ldots e_n)$ and n is the length of the English sentence.

Let us assume that the corpus consists of the pair of source and translated sentences, (f^k and e^k).

$f^k = (f_1^k, f_2^k, \ldots, f_m^k)$ where f_j^k is the j^{th} word in the k^{th} French sentence of length m

$e^k = (e_1^k, e_2^k, \ldots, e_n^k)$ where e_j^k is the j^{th} word in the k^{th} English sentence of length n

The parallel copora are available from Canadian parliamentary proceedings (the *Hansards*) and from Europarl data. Europarl data consists of proceedings from the European parliament, and consists of translations between several European languages

So, we let us assume certain things before we move forward. So, the task here is to translate a French sentence that contains a sequence of words. Remember the sequence right. It really takes you into the neural net (Refer time: 16:15) where we have utilized recurrent neural network for operating on the sequence of strings right. So, we going to be having a sequence of string that will describe a French sentence f and then f_j is the jth sentence whereas j equal to you can vary from 1 to m and m is the number of words in the French sentence ok. The same translated version could have the same number of words m words or more or less right.

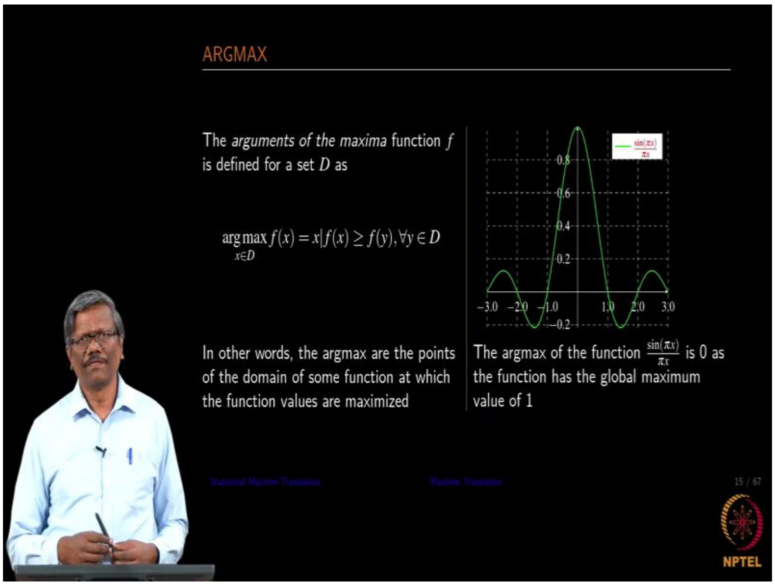
$$(f_1, f_2, f_3 \ldots\ldots f_m) \qquad f_j \;\; / \;\; j \in (1,2,3 \ldots n)$$

So, let us not use the same notation for English. The translated English sentence will be assumed to have a sequence e_1 to e_n where m is not equal to n in many cases and n is the length of the English sentence or the number of words in the English sentence ok.

So, when you look at the corpora there is not going to be just one sentence they are going to have many of them right. So, we have to define that ourselves. So, let us look at the parallel corpora that contain consists of the pair of source and translated sentences in this fashion f^k and e^k . We know that f represents the French sentence e represents the English sentence and k here refers to the kth sentence in that corpora. So, there could be about 10,000 or 20,000 versions or pairs of source and translated version k represents the kth sentence of both English and French ok. The parallel corpora are available from the

Canadian parliamentary proceedings you can also get it from there or you can get it from the Europarl data ok. So, Europarl is preferable because it contains lots of other languages store not just French and English.

(Refer Slide Time: 18:29)



ARGMAX

The *arguments of the maxima* function f is defined for a set D as

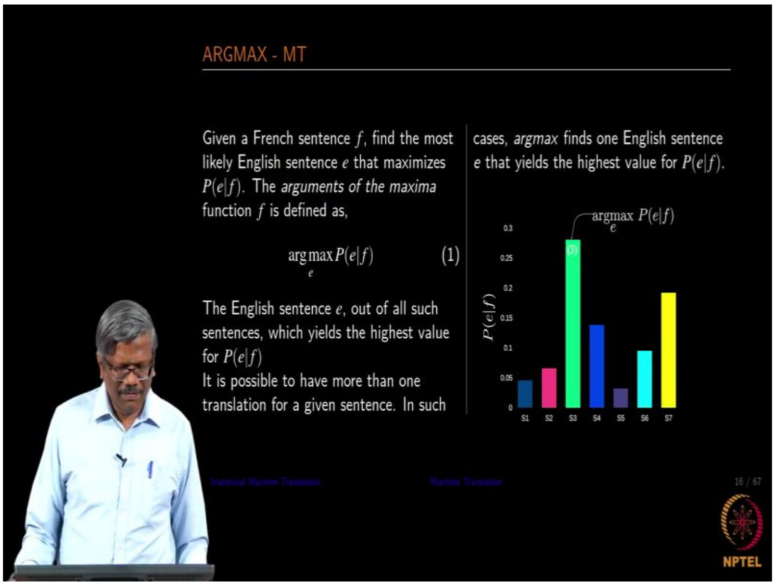
$$\arg\max_{x \in D} f(x) = x | f(x) \geq f(y), \forall y \in D$$

In other words, the argmax are the points of the domain of some function at which the function values are maximized

The argmax of the function $\frac{\sin(\pi x)}{\pi x}$ is 0 as the function has the global maximum value of 1

So, we spoke about this earlier, and then we also need to understand a few techniques that we would be using here. So, one is called the argmax ok. As I mentioned earlier there is going to be more than one sentence in the translated version for a source language right. So, which one should I pick up? So, that is what this is going to be telling us. So, if you have multiple sentences and then you find out what is the highest value or the probability and then pick that word. So, in this case we have the highest happening here and then the 0 index is your sentence that you want to pick up ok. So, if you have so many sentences arranged in this fashion and then you pick up that index where the maximum occurs.

(Refer Slide Time: 19:38)



ARGMAX - MT

Given a French sentence f, find the most likely English sentence e that maximizes $P(e|f)$. The *arguments of the maxima* function f is defined as,

$$\arg\max_{e} P(e|f) \tag{1}$$

The English sentence e, out of all such sentences, which yields the highest value for $P(e|f)$

It is possible to have more than one translation for a given sentence. In such cases, *argmax* finds one English sentence e that yields the highest value for $P(e|f)$.

Given a French sentence f, find the most likely English sentence that maximizes the probability of e given f $(P(\frac{e}{f}))$. It is the arguments this will give me the sentence here right. So, since as I mentioned there is a possibility of having more than one translation argmax will capture one English sentence that yields the highest value for the probability ok. So, this is the conditional probability and then the highest value for this conditional probability will be taken as the translated sentence.