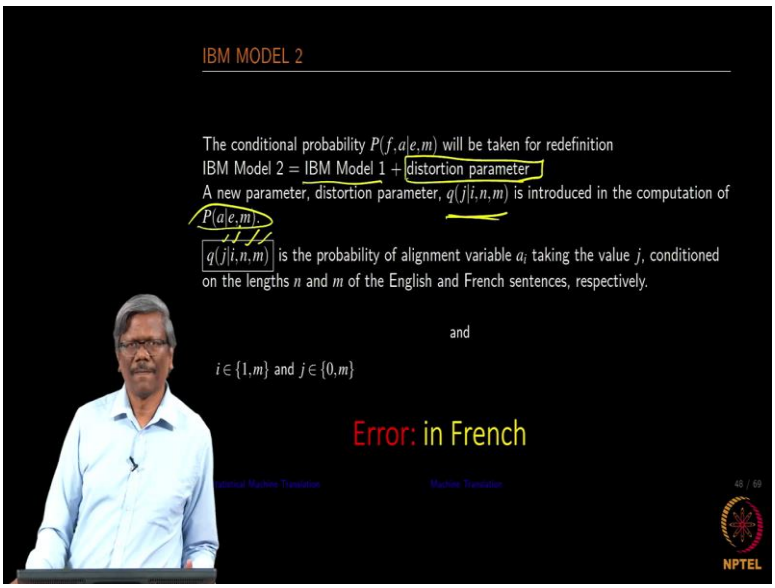
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture - 68
IBM Model 2

(Refer Slide Time: 00:15)



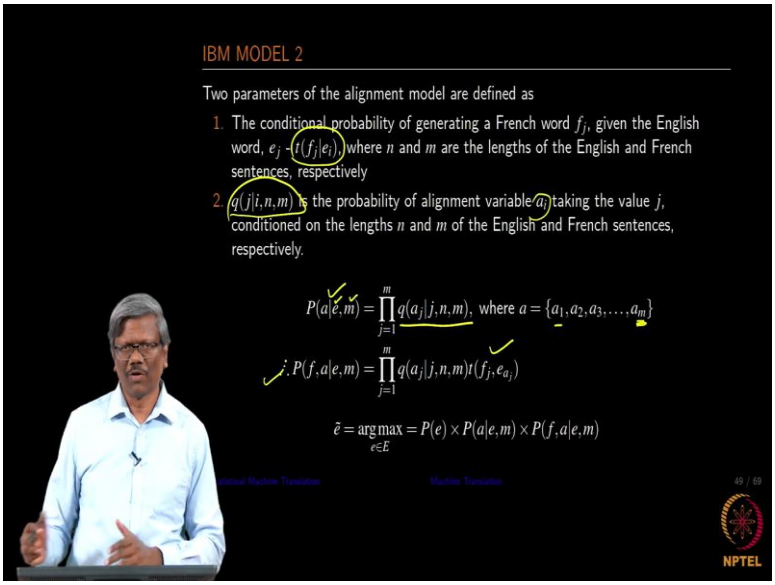
So, if you look at all the IBM Models, they are all based on manipulating the alignments most of the time ok. So, you will see in the IBM Model 2 also there is a slight chance that we are going to be bringing in which will make some more assumptions in terms of the alignments ok. So, this model uses whatever we saw in the IBM Model 1 and additional distortion parameter, we will describe what that distortion parameter is ok. So, this parameter is introduced in the computation of the translation model that we have described earlier, right.

So, earlier we had in this model in the IBM Model 1 there is two models that we found; one is the translation model based on the alignment the English sentence and the number of words in the French sentence, we were able to find the translation of the given word into French. And then another one is the alignment probability. So, in this Model 2 we will add an additional distortion parameter in the model ok. So, what that mean? You are going to be finding the alignment value j based on the number of French words that have

not changed; and then the number of English words in that sentence and then the I the word in English, ok.

So, this is the probability of alignment variable a I taking the value of j; conditioned on the length of n and m of the English and French sentences respectively ok, this is clear. So, this is one small change that is brought in to the IBM Model 2, ok.

(Refer Slide Time: 02:26)



So, here again we are going to be estimating two parameters; one is as we mentioned we still have to estimate the translation probability, given the English word find the French word ok. And then we need to estimate this probability of alignment for the variable i. So, here are the translation model contains the product of several probabilities of alignment. So, in this model we are going to be computing the alignment probability, a given the English sentence, and the number of French words, ok. So, this is given as the product of this distortion parameter that we have just described ok.

$$P\left(\frac{a}{e}, m\right) = \prod_{j=1}^{m} q((a_j \mid j,n,m) \text{ where } \{a_1, a_2, a_3, \ldots\ldots a_m\}$$

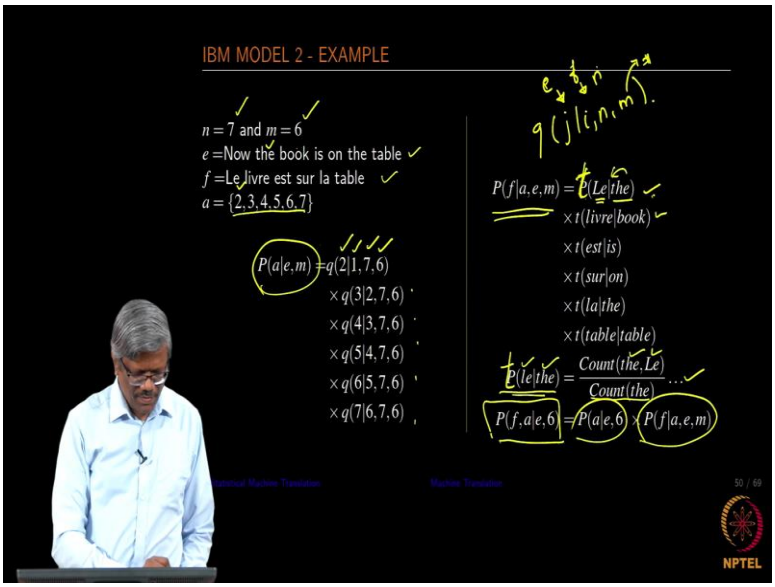
$$P\left(f, \frac{a}{e}, m\right) = \prod_{j=1}^{m} q((a_j \mid j,n,m) t\ (\ f_i,\ e_{aj}\ )$$

In this case a is a 1 to am; where m is the length of the French sentence. And then we have also the second one that we have to look at, right that is the translation probability. So, how do you combine? When you combine those to find the translation probability which came out from the noisy channel model; it is nothing but the product of the distortion parameter and the translation probability of the French word, given the English, so this should be this, given the English word at a_j, ok.

$$a_j = \arg\max_{e \in E} q(a|i,j,m) \times t(f_j|e,m) \text{ for } j = 1 \ldots m$$

And then finally, you want to find out the sentences using the decoder, this is your decoder ok. So, here it finds out the index of the sentences that we have generated using this model and then picks up the one which has the highest probability, ok. So, this is what arg max does for us, right. So, we have the language model, we have the translation model, and then I am sorry; we have the alignment model, and we have the translation model in this case, ok.

(Refer Slide Time: 04:53)



As in the previous case we will try to use an example to show how these alignment parameters can be computed and then move on from there, alright. So, again we consider the same set of sentences here, and then n =7 is the total number of words in the English sentence, and m = 6 is the number of words in the French sentence, and this is the

alignment variable that we have here. So, for each alignment variable we need to find the alignment probability, correct.

So, in this case we are taking the French word, right. So, in this case we are taking the alignment variable coming from the English sentence, right. And this is your word in the word that is aligned to the second word in English and so on, ok. So, let us write this, ok. So, this is the index for you in the English sentence, this is the index for you in the French sentence, this is the number of this is the English sentence, and this is the number of words in the French sentence, ok.

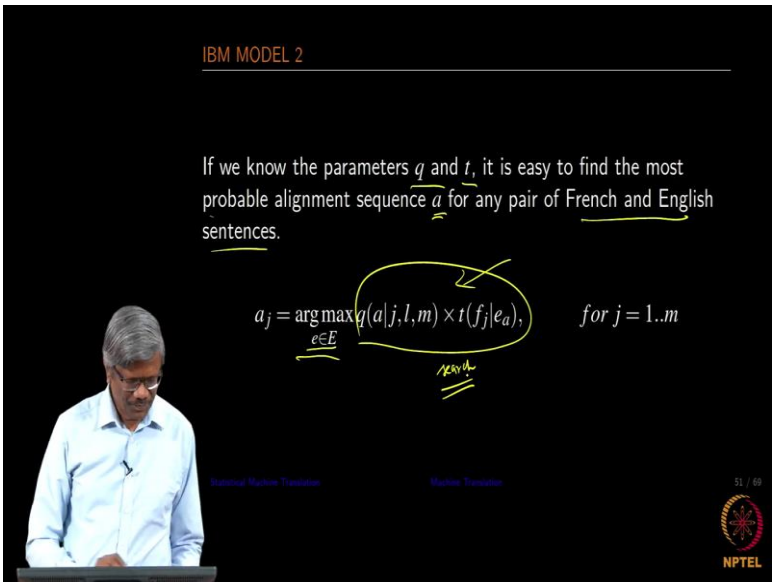
So, now let us go through this exercise to find out how the alignment probability can be computed ok. Let us look at the first one 2 OKs; the alignment probability for that variable can be computed using the second word that we are looking at here; and then this is the first word in French ok. So, this 1 is aligned to the second word in English; and then the number of words sorry this should be n the number of words in English I am sorry about that. Here we have the number of words in English and then here we have the number of words in French.

So, you take the second alignment. So, it is aligned to the second, if this is the second French word aligned to the third English word, right and so on. So, we can compute it in this fashion and then once you have the translation table that I had shown earlier a few slides back, you can look at the probability for each of the alignments here, right. Probability of a Le given the English word the ok, so in this way it should bet, let us call it as t ok; and then for each of this, we are getting the translation probability.

So, getting the translation probability for each one of the words and then multiply them and then for each one of them you need to get the count of the word, right. So, this again is, if you want to get this one, number of times Le and the are aligned together ok; and then you divide by the number of times the appeared in the entire corpus, ok. So, when we do that, you get the translation probability for Le given, the.

So, again when you want to finally, capture the translation probability of these noisy channel models ok. So, this is for the entire sentence that we are looking at, it will be equal to the alignment probability and the translational probability that we estimated.

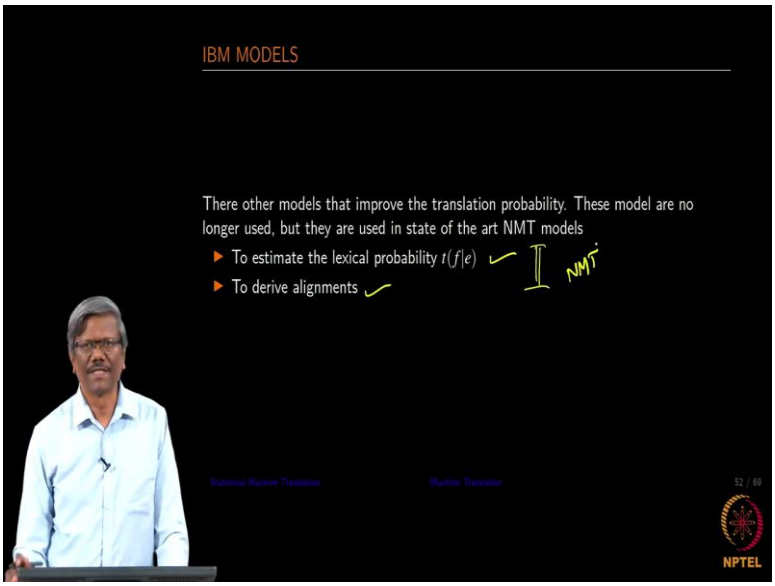
(Refer Slide Time: 09:18)



So, this is how you compute those values and then the best part of this is finding the alignment probability, right. So, that is the best part of the IBM Model, as I mentioned earlier the newer translation models do not use IBM Model as it is to do the translation. But they utilize the alignment, the computation of the alignment probability and the translation of translation probability of the French word given the English word, those models are used in the newer systems, ok.

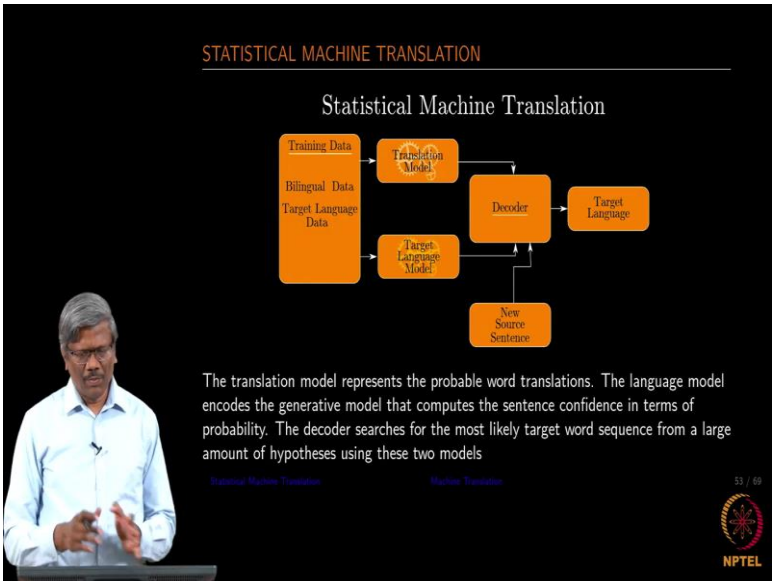
So, if you know the parameters of q and t, it is easy to find the most probable alignment sequence for any pair of English and French sentences, ok. So, again this is going to be a huge collection of alignment that you will see; and arg max will pick up which one has the highest probability and give you that. So, again it is a search mechanism to find out the right alignment, alright ok.

(Refer Slide Time: 10:36)



So, as I mentioned earlier there is another model that is available in the. So, we have only looked at Model 1 and 2; the other models are actually you know the improvements in Model 2 and so on. So, we do not want to really look at that models, these models really gave us a few things; one is how to estimate the lexical probability and then the second one is how to derive the alignments. So, these are the two important things that we will be used seeing in the neural net machine translations, alright ok.

(Refer Slide Time: 11:22)

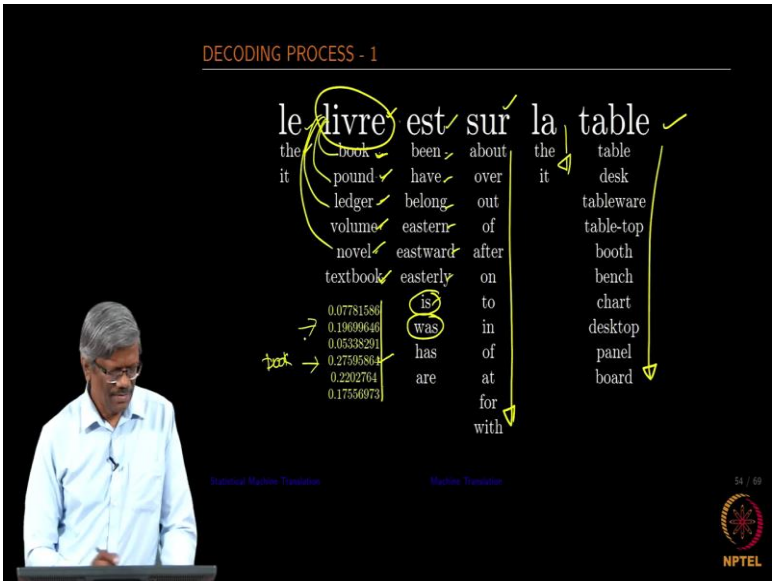


If you want to build a system around this, you know this is how you will do. So, what you will have is, you have the training data; the training data is nothing, but the bilingual data that you have the collection of all the French and English sentences. And then you have the target language data; for example, in this case from French to English we are going to do the translation.

So, we want to be able to have the problem the language model for us, for that you need the target language data which is English in this case. So, you have a huge collection of target language data, build your language model, get the bilingual data, get the translation model, ok. So, both are independent. So, we can do them in parallel; and we have the decoder model that utilizes both translation and the target model to finally, give the sentence that you are looking at ok.

And then once the models are developed, the decoder can take any new source sentence and then creator or provider target language sentence,. So, in this case again as I mentioned, this does a huge source operation; we will see how that search can be optimized in future sessions, alright.

(Refer Slide Time: 12:56)



So, to give an idea of how the decoding happens let us look at one small example. Let us say that, we want to translate this sentence into English, right; I used some dictionaries to get all these words. For example the le is mapped to, the as well as to, it; and then if you

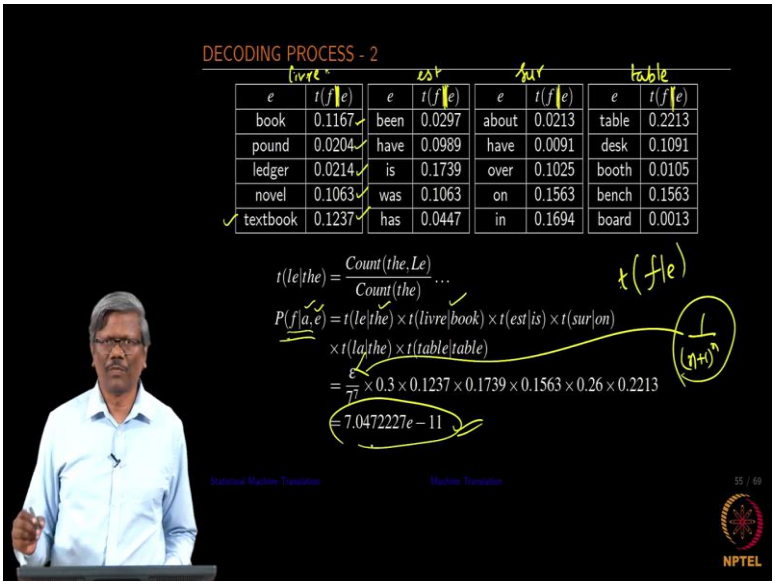
look at the French word here there are several translations possible for this, ok. So, we have book, pound, ledger, a volume, novel, textbook.

And then for this, you have variations, right; most of the time this or this would be aligned here. And then for this word we have so many English equivalents, again here; and then for the table again, there are several translations possible. So, what happens is, if you have a huge collection and then this is used, this word is used in several contexts, the right could be a book, it could be a ledger, it could be a novel, or it could be a textbook and so on.

The context varies every time since we are not really looking at the context; it is possible that there will be some alignment that would have happened between this and this, I am sorry, right. So, again you estimate the probability and then find out which one has the highest one and pick that up. In this case, I guess, I do not know this there could be this one, maybe this is aligned to this is equal to book in that in some order.

Again you will have several values associated with this, it is a probability distribution associated with this, for this translation and you pick the highest one from there ok. Since again I will mention that we are not using any contact information, this is only a lexical translation from word to word based on the learning's that we had through the data, ok. So, some translation might be useless as well in this case, ok.

(Refer Slide Time: 15:56)

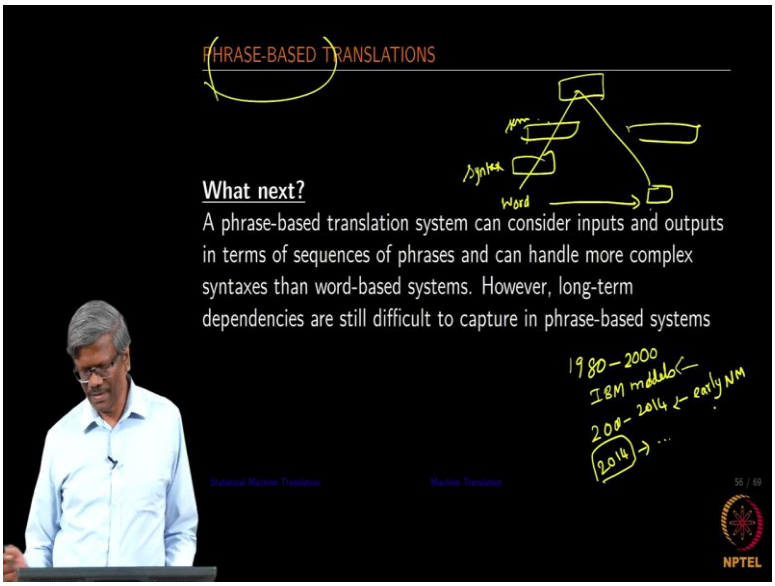


The second process is looking at the translation probability this is just one, ok. So, what we have is, we have the. So, we have tables partition in this fashion for each one of the French words ok; and then we have the translation probability that we have computed, it is completely based on them. And then based on them, probably you know the book would have some alignment, the pound will have some smaller alignment in terms of the probability, and then ledger too will have a very small set of alignments, a novel may be better, and then textbook may be higher in this case.

So, if I translate this, I would rather pick the textbook as the alignment ok, in this case. And then finally, you compute the alignment rather the translation probability by using this number ok; this eta is a normalizing constant that you have. So, you get the translation probability for each word here, that is listed; and then find it by multiplying each one of them and then use this one, to finally get the translation, ok. So, given the alignment and the English sentence, you get the translation and this is what you finally, end up, ok.

So, in all these cases you will find a very small value at the end ok. And this is only for one small set of sentences based on the alignment that we have picked. So, there could be several of this and that is where the decoder comes into play, searches through the entire collection of translations and then picks up one which has the highest probability, ok. So, in that case, again there is no guarantee that it picks up the best translation, it only picks up the translation that is guaranteed by the probability, ok.

(Refer Slide Time: 19:06)



Then what next; so these two models gave us you know some ideas in terms of how to really start the translation, given the data that we have in the training examples in terms of the parallel corpus, correct. But they are not going to be quite useful, because they are lexical models, which we even saw in the earlier case or even in the first session; in the triangle I am sure you remember that right.

We start with the word to word and then we have the syntax and then we have semantics, and so on right. So, what we have done so far is, only this. So, we are now able to instead of using the dictionaries that are available, and doing it in the old fashion way; we use the collection of the corpus and then try to do the word to word translation using this, ok. But this is not really going to help us in terms of figuring out the phrases, right. So, we need to be able to really identify the phrases, because certain phrases are aligned to one single word in French or vice versa.

So, we should be able to use those phrase models, you know in terms of file doing the translation. The recent model, that we are going to be discussing later is based on the phrases ok. So, with this I conclude this session. So, in the next session we will talk about, what is a phrase-based translation, and then later jump into the neural net models to see how translations can be achieved using the neural networks.

Especially if you look at this from 1982 to early 2000, we had IBM Models doing the job, they were not very successful in doing the translation. And then in early 2000, I am

not sure when until 2014 these are the early neural net models. From 2014 till today we have gone way far ahead in the translation models. And we are able to do a very good amount of research on this. And, a good amount of research went into the neural net-based translation model mostly because, of the sequence processing of the neural net model.

And then from 2014 till today, we have a really very good neural net model that does a very excellent job in terms of translation, still, it is not perfect ok. So, we have very good models, and we have good ideas, and I think in the next 10 years or so, we might be able to really have a very good translation model available to us. And then you remember the dream of doing this, a person let us say in a small village in India would like to talk to somebody in Mexico regarding some cultivation issues; they can run, they can talk in their own language and they still get their language translated in the black box.

And the person when he is talking in Spanish from there would be able to understand what the person from the remote village in Tamil Nadu spoke. And then the Spanish would be converted through the black box and it will be delivered as Tamil to the person in the local village in Tamil Nadu here, ok. So, that is the thinking that is going in the minds of the translators right now. Hopefully in the next 10 to 15 years, we will achieve that.

So, we will see how these models are taking shape in the neural net model, in the next session.