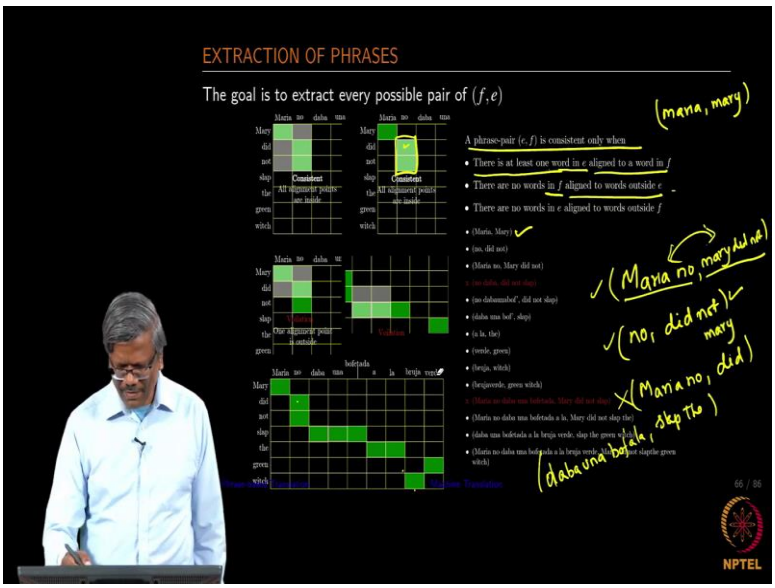
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture – 71
Extraction of Phrases

(Refer Slide Time: 00:15)



So, how do we now extract phrases? So, that is the next step right there could be many possible phrases that you can find in the alignment table that we have just created for phrases. So, how do we now get one after the other? For example, in this case a phrase pair; a phrase pair contains only a set of words that are very consistent, and if they follow the rules ok.

To be able to have a consistent phrase you need to follow these three rules that we have here ok. So, we will let us start for figuring out how many consistent phrases that we have and how many we can use for our translation purpose ok. So, the first one says there should be at least one word in e; aligned to a word in f. So, let us take the first one word for example, this one.

So, this rule works well right so, that could be a phrase, Maria, to Mary could be one of the phrases as I mentioned earlier. Here even one single word is considered a phrase ok. And, there are no words in f aligned to the words outside e. For example in this case there is none ok.

So, if you go back to the yeah here we have. So, you do not have anything around here. There is no alignment that you find in this below Maria; and then again in this case you do not find any alignment. So, this pair passed all three rules ok. So, this one so, this is a consistent phrase alright ok. Let us take another one, in this case so we have Maria did not right.

So, in Maria no, Mary did not Maria no. So, can this p another phrase? So, as I mentioned earlier please remember that the phrases that we are considering here are not linguistic phrases alright. So, in this case we are considering another phrase Maria no. So, according to the first one, there should be at least one word in e aligned to a word in f.

So, in this case Mary is aligned to this and did aligned to this and so on and so. So, this rule is satisfied there is more than one so, this rule is satisfied. The second one says that there are no words in f align to words outside e right. So now, this being your phrase we do not see any alignment points in this right, there is no alignment that we have seen.

So, the second rule is satisfied. Let us look at the third one; we do not see anything here. So, this one is consistent that is. So, these two phrases are consistent. So, we can say that this translation of Mary did not who yield this or vice versa got it. So, we have another phrase that is very consistent. Let us take a look at another one.

Let us take this alone right, that no is it consistent? Let us find out. So, there is at least one word yes in a align to a word in f. There are no words you do not find anything here, you do not find anything here right no alignment point that you find. So; that means, no and did not it is also another consistent alignment pair ok.

Let us take another one wire there is a violation ok. Let us take a look at this. So, if you take this as a phrase, the first one is satisfied there is at least one word that an e align to a word in f yes true. There are no words in f align to the word outside e ok. So, here this one is aligned to another one outside of this right.

So, this rule is violated so; that means, this phrase is not a consistent phrase right. So, this fashion we can go on and then created lots of phrases that are consistent and then use it for our translation purpose. So, this slap would be aligned to this word [FL] or [FL] I do not know Spanish so, pardon me ok. And, then ala is aligned to the.

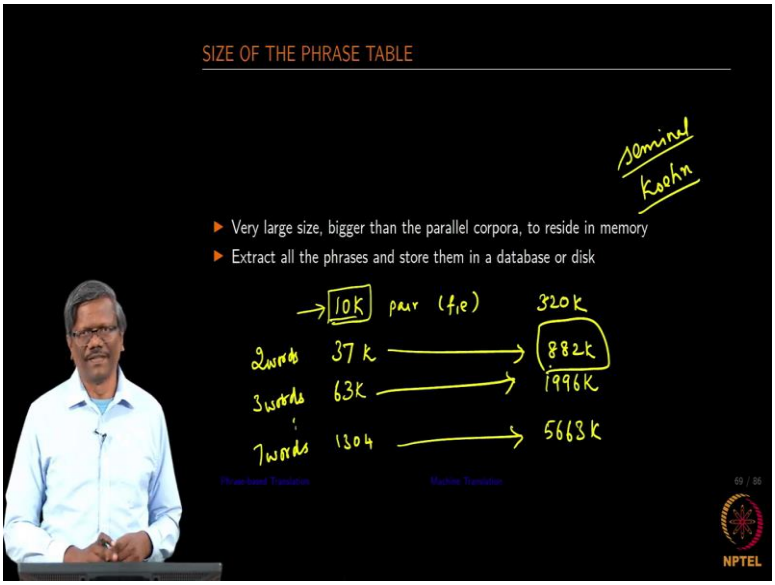
So, I cannot have alignment phrases and the right, because there is a violation here. So, in the same pair of fashion I cannot have something like this because, there is a violation here. So, can I have a phrase that encompasses all these words? So, in this case again if you look at rule 1, there should be at least one word in e aligned to a word in French yes or the foreign word that is true.

There are no words in f aligned towards the outside of e you know. And then again the third condition there is no violation. So, there is a possibility of having this as another phrase. So, that phrase would be the slap the would be aligned took. So, you see how the number of alignments or rather the number of phrases grows in terms of size right.

Again if you look at this, you can have another phrase consisting of a Greenwich right and then bruja Verde. So, even though they are interchanged the order is different here. So, when you do the translation we have to reorder in this. So, this what I was talking about earlier as well ok. This is the reordering of the phrases when you do the translation from Spanish to this to English ok.

So, this is possible. So, in this fashion, we can have all the alignments. So, can we have the entire sentences as one phrase? Yes, we can have the right alright. So, in this a slide, I have given various combination and this is not a complete list that I have ok. So, according to one of the research papers I am going to be showing some numbers for how many phrases that going to be creating by doing this? Let us come back to this little later.

(Refer Slide Time: 09:43)



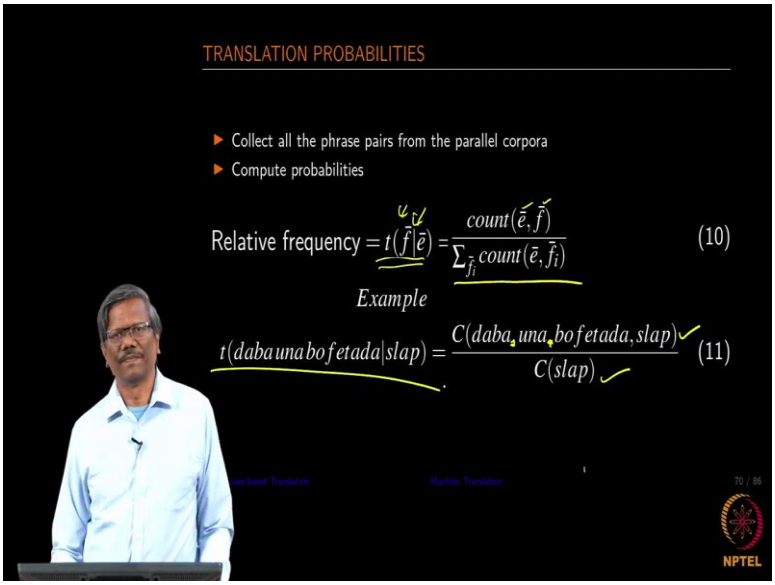
So, this size of the phrase table is going to be extremely large. I will just write down some of the numbers from one of the research papers that I have. And this is the seminal paper that you want to look at by Koehn ok. So, this is a very interesting paper and some of the slides that I have are based on this particular paper.

So, here if you have a corpus of letting us say 10 K pairs of words and if you have 2-word phrases, the number of phrases that you will have is about 37 K; and it is going to vary from a language pair to language pairs so, remember that. This just an approximate number that you want to keep in mind how this really grows in terms of the size; and then if you have 3 words the size is about 63 K ok.

And then if you want to have 7 words the size is about 130 K. And then since we are going to be having a very small pair of translation here. Or rather we not going to have a very small parallel corpus like this or the corpora size would be usually around 500 K. Let us take about 320 K as mentioned in the paper. And, then for 2 words it is about 882 K is the size of the phrases that you will find.

And, then for 3 words it is about 1996 K and then for 7 words it is 5663 K. So, look at the size it grows you know, how do we really store them. Can you keep all of this in the memory right for you to process? So, what normally you do is, you just store the pairs of phrases in the file system or in a database. And, then index them properly so that you can get them very quickly when you want to really do the translation process alright.

(Refer Slide Time: 12:27)



$$\text{Relative frequency} = t(\frac{\bar{f}}{\bar{e}}) = \frac{count(\bar{e}.\bar{f})}{\sum_{fi} count(\bar{e}.\bar{f})}$$

So, what are the translation probabilities that you can find? So, once you collect all the pairs now it is up to us to go and then find out the parameters or estimate the scores for yields of those. So, again as we had done earlier, we are going to be finding out the translation probability, it is a conditional probability given the English word.

What is probable? French; I am sorry, given the English phrase what is probable? French phrase so, for that you need to find out how many times these two are aligned and over the entire corpus so, this is one way. Or if you want to find out as an example ok; so, how many times this as happen divided by the number of times you found slap ok. So, this count will give you the probability score for you.