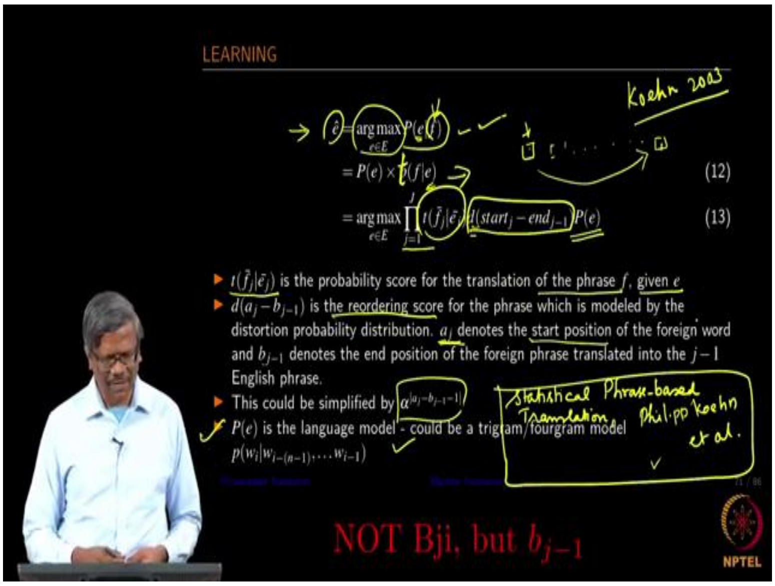
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture - 73
Learning/estimating the phrase probabilities using another Symmetrization example

(Refer Slide Time: 00:15)



LEARNING

Koehn 2003

$$\rightarrow \hat{e} = \arg\max_{e \in E} P(e|f)$$

$$= P(e) \times t(f|e) \quad (12)$$

$$= \arg\max_{e \in E} \prod_{j=1}^{J} t(\bar{f}_j|\bar{e}_j)\, d(start_j - end_{j-1})\, P(e) \quad (13)$$

- $t(\bar{f}_j|\bar{e}_j)$ is the probability score for the translation of the phrase f, given e
- $d(a_j - b_{j-1})$ is the reordering score for the phrase which is modeled by the distortion probability distribution. a_j denotes the start position of the foreign word and b_{j-1} denotes the end position of the foreign phrase translated into the $j-1$ English phrase.
- This could be simplified by $\alpha^{|a_j - b_{j-1} - 1|}$
- $P(e)$ is the language model - could be a trigram/fourgram model $p(w_i|w_{i-(n-1)}, \dots w_{i-1})$

Statistical Phrase-based Translation Philipp Koehn et al.

NOT Bji, but b_{j-1}

Once these phrases are found out, the probabilities are filled in the table, we should be in a position to learn the model, right. So, as we know that this is a search process, given the, French sentence. As we know very well that this is a search process and argmax will give you the index that has the maximum probability for the French sentence that going to be that is going to be translated into an English sentence.

And, then you pick up that sentence using a search operation and say that, this is how I am going to be providing you the translation for a given French sentence, right. So, we know well. So, when you use the Bayes rule, actually the translation is inverted that also we know and it is broken into two steps, one is the, a language model for English, another one is another probability which will give you another score, where when the word or the phrase in English is given.

And using that you will find out the probability for the phrase in French, right. So, and this is the parameter that you would estimate this should bet, this is the translation probability given e, alright. And then if you take the whole thing right, so, it is going to be the product of the translation probability of all the, phrases being translated into this. And then there is another one called a distortion parameter. So, that this tells you or it this is like a penalty parameter.

Where when the word that you are going to be translating from English to the foreign word right, so since it is inverted due to the Bayes rule we are doing this. And if the foreign word is far apart from the word where you are doing the translation at this point in time and then the number of words that are counted in between. And that particular number would be used to provide a penalty for the score that we are computing. The reason is if the words are not close enough we are actually giving a higher penalty.

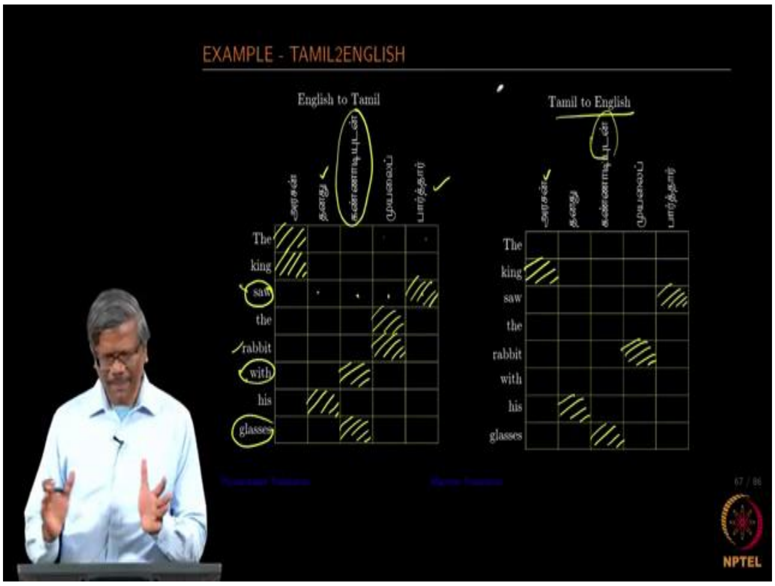
So, you want to be able to find out words that are closer to each other during the translation. So, here t f given, this translation probability is this score for the translation of the phrase f given e, right. And then this is sorry this should be a; this should be b. So, based on the paper, because I do not want you to get confused. So, with different notation, if you read the paper by, I think it I 2003, let me give you the name.

So, this paper is Statistical Phrase-based Translation ok. So, you search for this I think this is freely available. So, these notations are taken, from this paper. And then the distortion parameter also is simplified in this fashion it is an alpha that we have chosen, usually this alpha would be 2 or so. And then we know that P is the language model that could be a trigram or fourgram model, using which you can find I identify what is going to be the probability of the next word given, the first two words and so on ok.

So, this one is computed using a huge corpus of the English language. We could compute this from our phrases that we have obtained and then this parameter could be computed based on how far the distortion is going to affect our translation process ok. So, this helps you in terms of identifying the reordering score ok. So, where a j is this starting position of the foreign word and b j I denote the end position of the foreign word translated into j minus 1 English phrase ok.

So, how does learning happen here? So, in this case you know it is a product of all these three that we have mentioned, right or in order to get out of that underflow problem, what you can do is you can use a log. So, it becomes an addition for us and then you can move on from there ok, alright.

(Refer Slide Time: 06:45)



EXAMPLE - TAMIL2ENGLISH

So, as I mentioned earlier let us look at this particular language, right so, unless you take a different language pair it is not going to pause it is not going to be easy for you to understand the alignment of phrases part. So here I have taken a foreign language, it is not a foreign language to me, this is my mother tongue and this is English. I have a sentence, “the king saw the rabbit with his glasses”, right. So, even though this sentence is a very confusing one I have taken this and then the equivalent translation in Tamil is [FL].

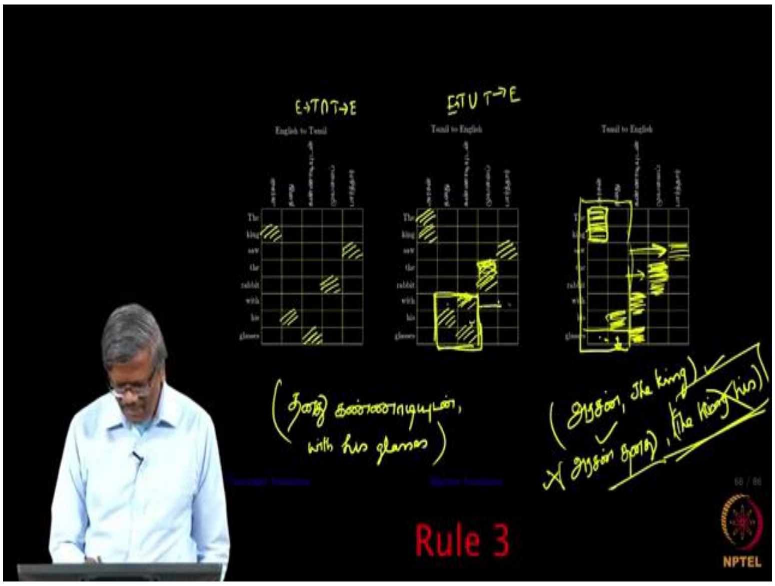
So, “the king saw the rabbit with his glasses”. So, what we need to do is, we need to first fill the alignment table. So, I am going to be filling the first one English to Tamil ok, right and then, “his” is translated into this world and then this a very interesting thing in Tamil, right. So, you do not have a separate word like in English, “with glasses” is translated into this [FL] ok. So, you will have alignments in this case here and here ok.

And then “the rabbit” is the [FL] here and then “saw” is coming towards the ends. You remember the reordering problem, right, so now, if you look at this word, saw, if you want to translate this into this, how many words if you are here, you have to skip one, two, three, four, there are four words from here you may have to look at. So, the penalty that we are seeing there would be alpha that you are having would be equal to 4 huge penalties for having as separation in this, right.

So, this is the trouble that you have to reorder the sentences when you translate from the foreign language to the English language. So, again then let us go to the Tamil to English translation. So, straight away it is [FL] to "the king" and then "his" is [FL], "glasses" is translated into a [FL], but we cannot have an equivalent of this and you have to do a lot more linguistic processing to really break this into this so that it becomes another a word in Tamil. But, this is the standard translation so, we leave it here.

And then "rabbit" or [FL] is equivalent to the rabbit and then, "saw" here, sorry it should be here, right. So, this is the English to Tamil and then Tamil to English. Now, we have to symmetrize this, so, what do we do for this symmetrization process? One is to take the intersection of this first and then create on a table of this type and then use align use a union and create another alignment table. And then use the rules to form the consistent phrases ok.

(Refer Slide Time: 11:01)



English to Tamil

Tamil to English

Tamil to English

Rule 3

So, in this case I am going to be copying from what I had earlier when you do the intersection, so, this is the intersection process. So, what are the common word that you will find "king", right? So, you just mark "king" sorry and then "saw" and then "rabbit", right. I think we covered everything. So, this is the place where you start to grow your phrases. And then when you come here to the union part so, you will cover this, it will be the same as what you have done here. And you will have this where is the "the rabbit" alright, so, this is our union part, so, when you start to grow this in the next one. So, we need to look at this, so, you will fill this up and then you will also fill this up.

So, it will be where equivalent to your union part only in this case. So, I quickly fill all of those. So, now, these are the phrases that we want to look at. So, this contains all our phrases we may want to take a look at. So, in this case what we can do is. So, this one is one phrase rights say for example, this is a valid one because you do not find any foreign word-aligned or any English word-aligned according to the rule 2 and 3 and there is at least one word that is aligned here, right.

So, this role gives us this as our alignment. So, can we have this as one alignment? Let me erase this. So, I am going to be looking at this, is it possible to have an alignment in this case? So, where we have; is it a legal alignment? No, right, there are alignment points here, right? So, this is not a legal alignment, so, I cannot have this. So, according to rule 1 we have many points that are aligned.

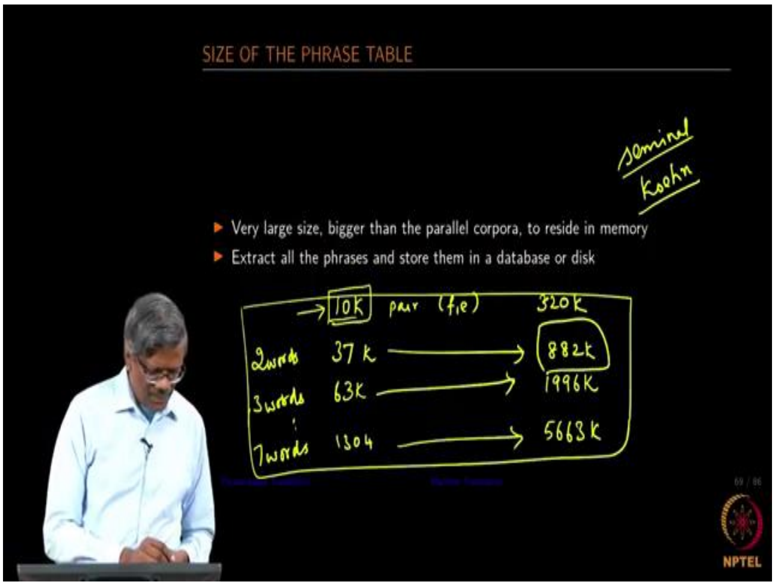
So, we can say that rule 1 is satisfied, the rule 2 says and rule 2 and 3 says there should not be any alignment along the foreign direction and there should not be any alignment along with this. So, in this case there is one rule that is satisfied because you do not find any alignment points here. But, the rule that we have will not let you align this phrase. So, we do not have a good alignment here.

And also it does not make any sense even in the, language Tamil. Or, in rather in Tamil it makes sense, but in English it does not make any sense I am sorry ok. So, we have discarded this one. So, in the other case for example, with his glasses. So, let us take this, so, in Tamil it becomes, so, the other one is with, so, let us look at whether this phrase is a consistent phrase and we can use it.

So, rule 1 applies right, there is nothing around here. Rule 2, I am sorry, rule 1 says there should be at least one alignment point for the foreign language and it is true. And then rule 2 says, there should be no alignment point along this direction and no alignment point in this direction too, right. So, according to the rules this is a consistent phrase, right. So, in the same fashion, so we can look at this phrase and then say that this is another phrase that is and this is another pair of phrases that is consistent, so we can use this.

So, in this fashion you keep creating more and more phrases and then fill your translation probability table using the translation that you find here using your phrases by doing the symmetrization process, alright.

(Refer Slide Time: 19:04)



SIZE OF THE PHRASE TABLE

seminal Koehn

- Very large size, bigger than the parallel corpora, to reside in memory
- Extract all the phrases and store them in a database or disk

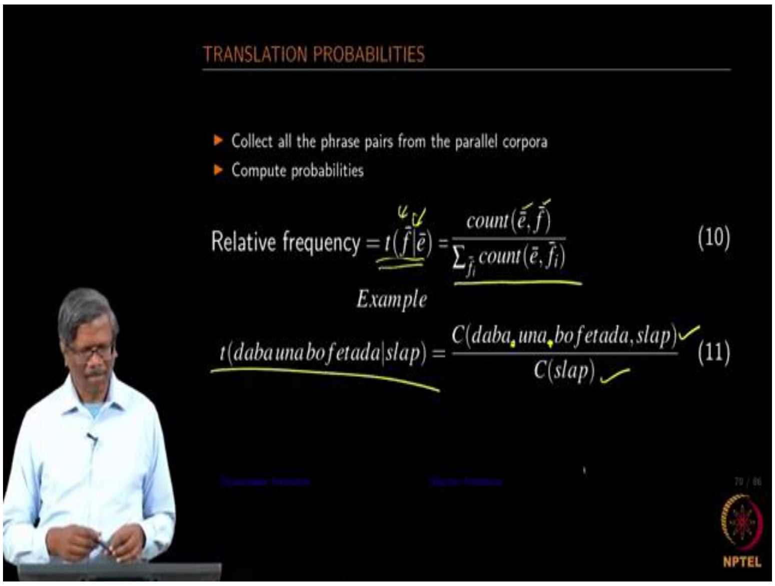
→	$\boxed{10K}$ pair (f,e)	320K
2words	37 k →	$\boxed{882k}$
3words	63K →	1996K
⋮		
7words	1304 →	5663 K

69 / 86

NPTEL

So, as I spoke earlier the size of the, the translation table is very huge. So, we may want to store it in the disk or in the file system or in the database somewhere. So, that you can quickly retrieve whenever you need for the translation process alright.

(Refer Slide Time: 19:23)



TRANSLATION PROBABILITIES

- Collect all the phrase pairs from the parallel corpora
- Compute probabilities

$$\text{Relative frequency} = \underline{t(\overset{4}{\bar{f}}|\overset{✓}{\bar{e}})} = \frac{count(\bar{e},\bar{f})}{\sum_{\bar{f}_i} count(\bar{e},\bar{f}_i)} \tag{10}$$

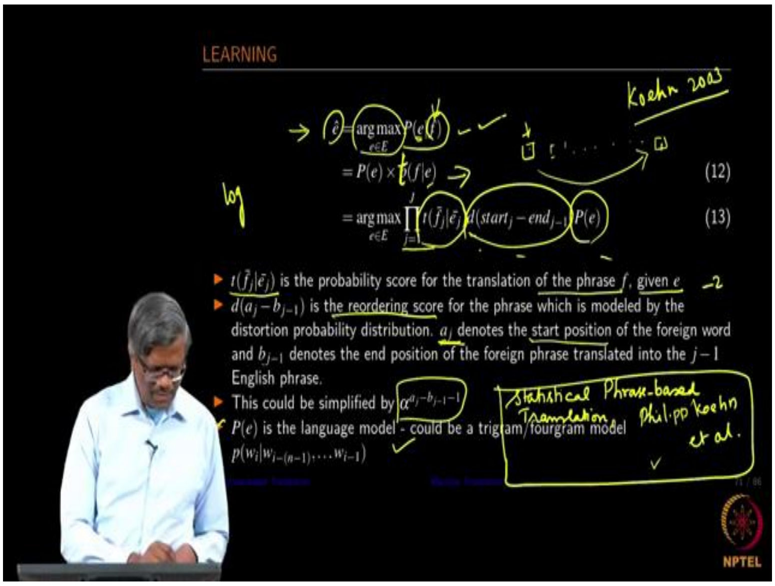
$$Example$$

$$\underline{t(daba\,una\,bofetada|slap)} = \frac{C(daba,una,bofetada,slap)✓}{C(slap)✓} \tag{11}$$

70 / 86

NPTEL

(Refer Slide Time: 19:27)



LEARNING

$$\hat{e} = \arg\max_{e \in E} P(e|f)$$

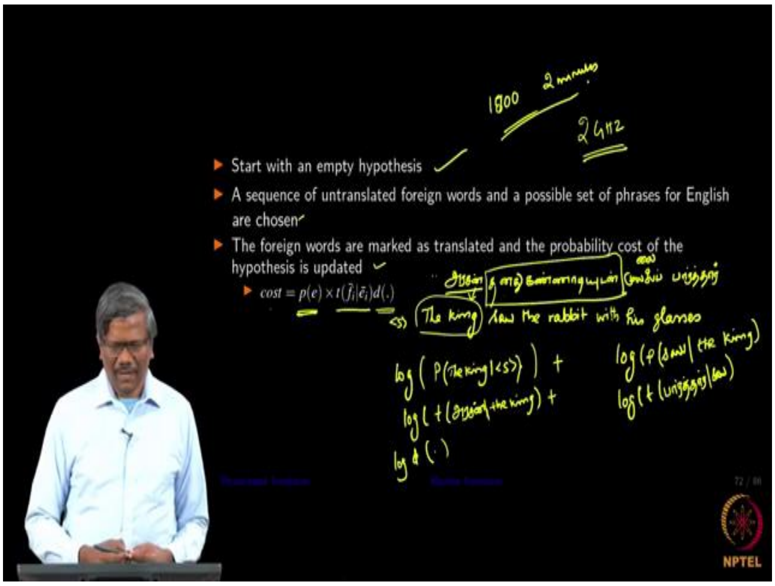
$$= P(e) \times p(f|e) \quad (12)$$

$$= \arg\max_{e \in E} \prod_{j=1}^{J} t(\bar{f}_j|\bar{e}_j)\, d(start_j - end_{j-1})\, P(e) \quad (13)$$

- $t(\bar{f}_j|\bar{e}_j)$ is the probability score for the translation of the phrase f, given e
- $d(a_j - b_{j-1})$ is the reordering score for the phrase which is modeled by the distortion probability distribution. a_j denotes the start position of the foreign word and b_{j-1} denotes the end position of the foreign phrase translated into the $j-1$ English phrase.
- This could be simplified by $\alpha^{a_j - b_{j-1} - 1}$
- $P(e)$ is the language model - could be a trigram/fourgram model $p(w_i|w_{i-(n-1)}, \ldots w_{i-1})$

So, we spoke about the learning of this using the Noisy-Channel model as well as the, Bayes rule. And each of these parameters could be estimated from the data that we have for the pair of for every pair of sentences you find in the parallel corpus. And, then once you have created the tables corresponding to all the translation probabilities you are ready to decode the new foreign sentence alright.

(Refer Slide Time: 20:06)



- Start with an empty hypothesis
- A sequence of untranslated foreign words and a possible set of phrases for English are chosen
- The foreign words are marked as translated and the probability cost of the hypothesis is updated
  - $cost = p(e) \times t(\bar{f}_i|\bar{e}_i)d(.)$

1800 2 minutes
2 GHz
அரசன் ஒரு கண்ணாடியுடன் முயலைப் பார்த்தார்
<s> The king saw the rabbit with his glasses
log (P(The king|<s>)) +
log (t (அரசன்|the king)) +
log d (.)
log (p (saw| the king)
log (t (பார்த்தார்|saw)
NPTEL

So, in this case how do we do it during the learning process? We start with an empty hypothesis, sequence of untranslated foreign words, and a possible set of phrases for

English, chosen foreign words are marked translated and the probable cost of the hypothesis is updated. So, how do you do that use using the product of the language model, the translation probability of these phrases and the, a distortion value or the penalty that you have ok? So, the cost is computed for every possible translation property.

$$\bar{e} \;=\; argmax\, P(\frac{e}{f})$$

So, it becomes a huge table for you to have a lookup when you finally do the decoding process. So, how do we take it forward? Let me use some examples here. We take the pair of sentences, I am just thinking about whether you should take the Spanish one or the Tamil one. Let me take the Tamil one for you ok. So, where we have, I am used to writing it in the word script, those who do not follow the word script it is, that is right. So, when you start the process you take the first word we know that it is aligned to this right.

So, when you take this as the starting word, so, what happens is since we are going to be doing the language model for English, we do not consider this you only consider this, right. So, we have a starting symbol and then we need to find out “the king” given the starting symbol.

So, we have created this already. So, we go and look at the table and then find out what is the probability of this, right. So, here let us say that I am going to be using the log of this and then the second one is again the translation probability of given the English word what is the foreign phrase that you will find. And, then you have the distance parameter, in this case, it is in the same place, so, it is not going to be a very high penalty.

Again that is a log value for you right, so, this is for one of these. And then what you do is you take the next one, the next word here and then find out the probability of this in the same fashion we have found out in this case what is going to happen for the language model, what is the probability of finding the next word saw given “the king”, right. And then we have another log for the translation probability.

So, in this case I am sorry should be, so, we keep doing this for all these possible phrases, that becomes a huge table by itself right for you. So, for every possible combination of the phrases that you find in a given sentence you do this. You take this and then sometimes this goes as an individual phrase and then sometimes the words, the two words will go as a phrase.

So, you keep doing this the combinations of all the phrases that you will find or that you have computed earlier and then finally, arrive at a, huge list of translations for the given foreign language. That means, you are going to have a huge list of English sentences that you want to pick up and then use the argmax to pick the right one at the end ok. So, during this training process you keep filling this table and then that is what would be available for the decoder.

And then a new sentence is given in the foreign language, it starts looking at each word. And then trying to find out what is the probability of translating that particular word or phrase into the English language. And then start building that table and then finally, the decoder will find a lot of sentences and then using the argmax pick up the best sentence and then provide that as the translated version for you. So, I have seen in some studies using a two GigaHertz machine it is possible to build close to 1800 or 2000 translation in about 2 minutes ok.

So, now that we have a lot more power and then if you use some parallel-processing capabilities in this wherever possible, you should be able to do the translation and learn faster. So, with this I close the statistical-based translations. So, in the next session we will see neural net-based translation models using sequence to sequence, recurrent neural network either using LSDM or GRU or let us see if we can find different models.