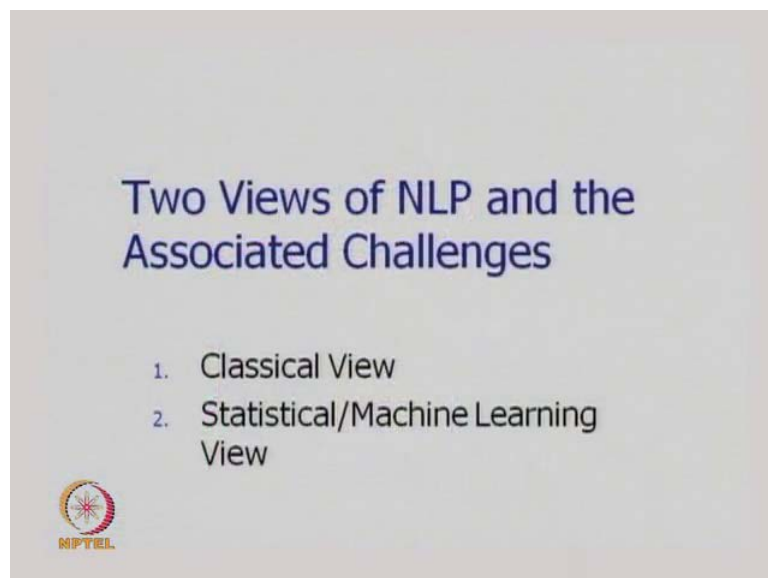


Natural Language Processing
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 6
Noisy Channel: Argmax Based Computation

We continue our lecture on natural language processing. Today's lecture is on noisy channel modeling, it is used in natural language processing and a very important technique called argmax computation, let us proceed.

(Refer Slide Time: 00:36)




Now, we have mentioned many times that there are two views of natural language processing, and there are associated challenges, the classical view and the statistical machine learning point of view. In classical view, knowledge and rules are used which are created by human beings lexicographers, linguists they capture the regularities of a language and they introduce rules which are used by the machine. In statistical machine learning view, what happens is that the data is processed by the machine and it generates the rules through machine learning techniques, it also generates the probability values.

(Refer Slide Time: 01:25)

Example of Sentence labeling: Parsing

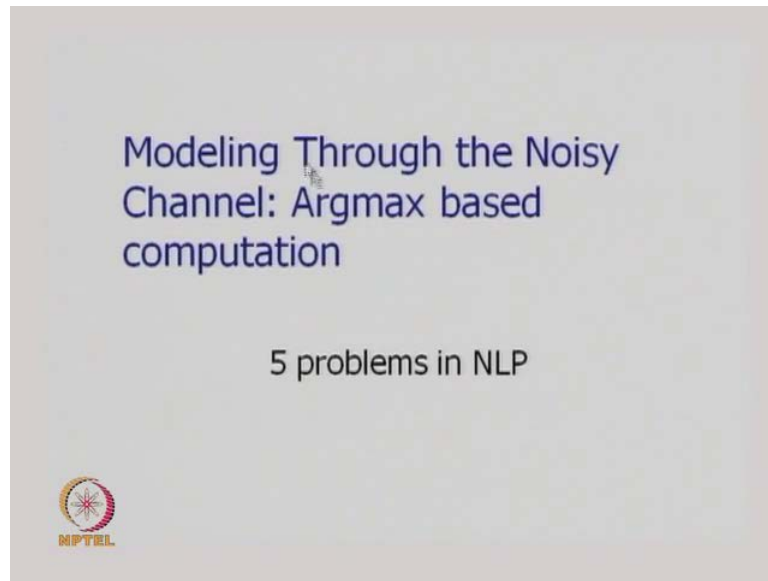
```
[S1[S[S[VP[VB Come]][NP[LNNP July]]]]  
[, ,]  
[CC and]  
[S [NP [DT the] [JJ IIT] [NN campus]]  
[VP [AUX is]  
[ADJP [JJ abuzz]  
[PP [IN with]  
[NP [ADJP [JJ new] [CC and] [ VBG returning]]  
[NNS students]]]]]]]  
[. ]]
```



In the last class, we had these sentence labeling problem in the form of parsing out of all the sequence labeling tasks we have said the parsing is one situation where the labeling problem is not so obvious. So, the sentence is shown here as come July and the IIT campus is abuzz with new and returning students, so this whole sentence has been given a bracketed structure which in the last class we saw is equivalent to a tree. So, a two dimensional tree is equivalent to a one dimensional bracketed structure of sentence where the brackets capture the trees and the sub trees.

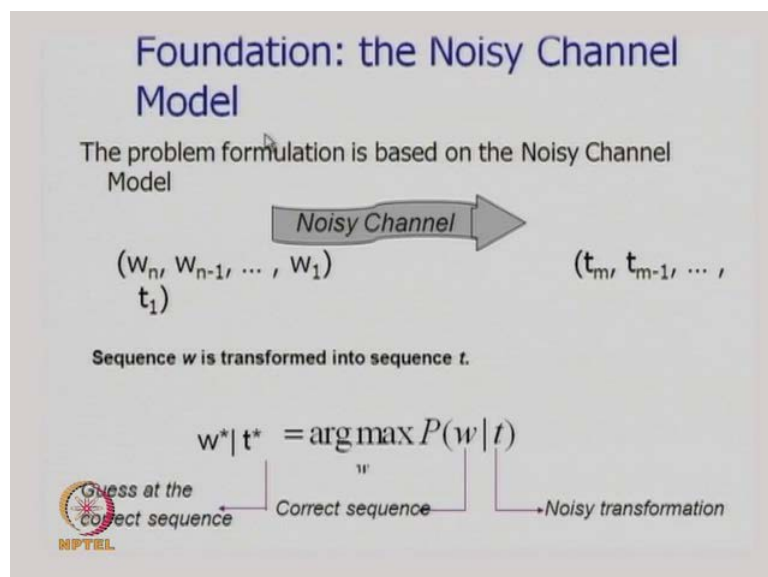
So, the IIT campus for example, IIT and campus they together form a unit, then it is a determiner coming before IIT campus, so the IIT campus, these three together form a noun phrase. So, you can see that there is this n p note which has the children as the IIT and campus, so these three units together form the n p tree. So, the whole structure the whole structure is actually a tree with sub trees within tree and therefore these becomes completely a sequence labeled data.

(Refer Slide Time: 00:34)



Proceeding, we now introduce modeling through the noisy channel argmax based computation, so the question we are asking is we understand natural language data is in the form of raw text and on top of that we produce labels. So, it could be part of speech tag labels named entity labels or parse trees which are bracketed structures. So, the question now is how do we produce these labels these labels are produced by a technique which is mentioned in the slide modeling through the noisy channel argmax based computation. So, we will illustrate this technique by means of five problems in natural language processing.

(Refer Slide Time: 04:05)



The picture here shows the foundation of the noisy channel model the problem formulation is based on the noisy channel model idea which is borrowed from speech it dates back to 1960s. The noisy channel was a metaphor for any computation where the input had to be transformed into an output and there were noise coming at every stage of processing. So, an example of that would be a automatic speech recognition where the speech signal having been heard is converted to a piece of text in in the form of written form.

So, a speech signal converted to written text form, so this is a transformation which happens over a noisy channel because on the way from speech signal to text there can be different sources of error. This is modeled by the noisy channel, so this this technique came from probability and speech, this has been erupted to natural language processing and we will see very soon how the technique comes to great use in natural language processing problems.

If you look at the figure here, this is the noisy channel on the input site of the channel there is a sequence $w_1 w_2 \dots w_{n-1} u$ to $w_1 w_2 \dots w_{n-1} w_n$ up to w_1 . This is a sequence which becomes another sequence $t_1 t_2 \dots t_{m-1} t_m$, so entities w_1 to w_n go over the noisy channel, another set of entities t_1 up to t_m . So, sequence w is transformed into the sequence t now here one could think of finding the best possible w^* given t or one could find out best possible t^* given w . We will make this little more precise, but please note the use of a function here called argmax, let us understand this argmax function, I will now write about argmax computation.

(Refer Slide Time: 06:47)

$$y = f(x)$$
$$f(x)^* = y^* = \max(y)$$
$$= \max(f(x))$$

Maximum value of $y = f(x)$

So, let us say we have y equal to $f(x)$ what is the difference between y^* equal to \max of y is equal to \max of $f(x)$, so this is equal to f^* equal to y^* . So, f^* equal to y^* equal to \max of y is equal to \max of $f(x)$, so this corresponds to the maximum value of y equal to $f(x)$.

(Refer Slide Time: 07:31)

Compare "max" with
"argmax"

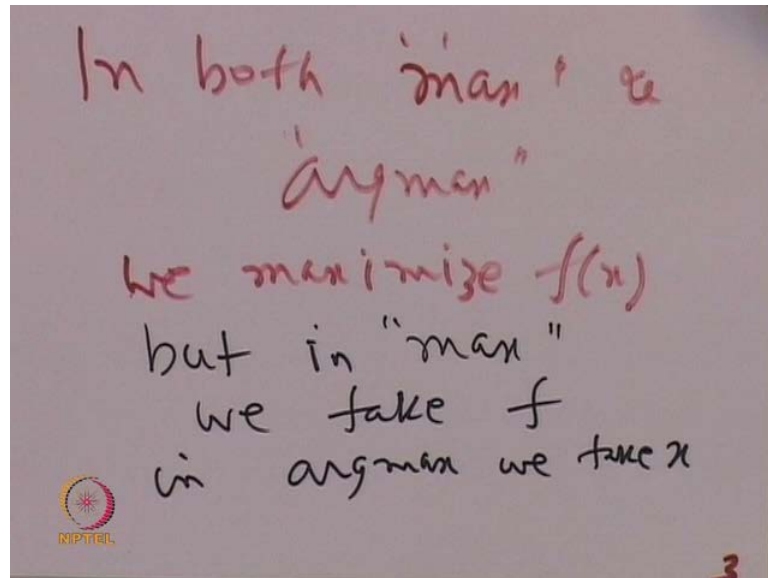
$$x^* = \arg\max_x (f(x))$$

Find that x which
maximizes $f(x)$

Now, if we compare this with $\arg\max$, so compare \max with $\arg\max$, now we will write x^* equal to $\arg\max_x f(x)$ over all x . So, it means that find out that value of x find that x

which maximizes $f(x)$, so while comparing max with argmax we find that value x^* which maximizes the $f(x)$ value.

(Refer Slide Time: 08:23)



So, the thing to note is that in both cases in both max and argmax in both max and argmax we maximize $f(x)$. So, in both max and argmax we maximize $f(x)$, but in max we take f in argmax we take x , so both max and argmax what is maximized is $f(x)$. There is no change with respect to that both argmax and max concentrate on maximizing $f(x)$ max takes the f value argmax takes the x value. So, find out that value of x for which $f(x)$ is maximum and return that, so that is the meaning of argmax whereas, max is maximize $f(x)$ and return f that is the difference so argmax computation is central to many of the things we will discuss now.

(Refer Slide Time: 09:40)

Bayesian Decision Theory


- Bayes Theorem : Given the random variables A and B,

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

$P(A|B)$ Posterior probability

$P(A)$ Prior probability

$P(B|A)$ Likelihood



We see the transparencies and A in this case we are mentioning Bayesian decision theory, so argmax computation noisy channel modeling all these are essentially statistical processes for doing A task and the heart of this is Bayesian decision theory. So, Bayes theorem is A very well known theorem of probability, it says that given the random variables A and B probability of A condition on B that is probability of A given B is equal to probability of A into probability of B given A divided by probability of b.

So, this is the Bayesian theorem and it is A very celebrated theorem due to Bayes who was A seventeenth or eighteen century A mathematician and the this very simple formula has found lot of application in decision theory. Now, this can this theorem can be very easily proved if you take P B on this side, then you have P B into P A given B equal to P A into P B given a. So, is this true this is true because the left hand side now becomes P A dot B P A and B and the right hand side is also P A and B or P B and a, now since it is commutative P A is equal to P B and A and therefore, this theorem is true.

So, proof of this theorem is very simple, you just have to use the fact that P B into P A given B is nothing but P B and A and this side is P A and B. So, this is a Bayes theorem now there are some terms which are used in discussing Bayes theorem P A given B is called the posterior probability, so how does the probability of A change given A the knowledge of B. So, you can contrast the posterior probability with the prior probability P A is the prior probability P A given B is the posterior probability. So, we are given the

prior probability of A, now the knowledge of B has arrived, so how does the probability of A change.

So, this is modeled by probability of A given B this is posterior probability this is prior probability as I have already mentioned and probability of B given A is the is called the likelihood probability how likely it is that B occurs given A. Now, A I wonder as to what difference it makes whether we work with P A given B or P B given A, so this is the way a valid question should be worked with P A given B or P B given A, now this is a very valid question. We will see that when we take up the particular problem that decides which probability we should work with which probability is better to work with which is more convenient is it P A given B or P B given A we will take some examples.


(Refer Slide Time: 13:32)

Bayes Theorem Derivation

$$P(A \cap B) = P(B \cap A)$$

Commutativity of "intersection"

$$P(A) P(B | A) = P(B) P(A | B)$$
$$\Rightarrow P(A | B) = \frac{P(A)P(B | A)}{P(B)}$$



Now, Bayes theorem derivation it has been already mentioned that P A and B is equal to P B and A and from that the theorem follows.

(Refer Slide Time: 13:41)

To understand when and why to apply Bayes Theorem


An example: *it is known that in a population, 1 in 50000 has meningitis and 1 in 20 has stiff neck. It is also observed that 50% of the meningitis patients have stiff neck.*

A doctor observes that a patient has stiff neck. What is the probability that the patient has meningitis?

(Mitchel, Machine Learning, 1997)

Ans: We need to find

$P(m | s)$: probability of meningitis given the stiff neck



Now, we would like to understand when and why do I apply Bayes theorem which is equivalent to asking should we work with $P(A | B)$ or $P(B | A)$. Here is an example I describe this example is from tom Mitchel's celebrated book machine learning 1997, this is the example it is known that in a population 1 in 50,000 has meningitis. Meningitis is a disease of the brain sometimes very fatal and 1 in 20 has stiff neck, we know that 1 in 50,000 has meningitis that is a rare phenomenon where as stiff neck is a much more frequent occurrence. Therefore, 1 in 20 has stiff neck it is also observed that fifty percent of the meningitis patients also have stiff neck.


So, this is the case that in case of meningitis about half the patients suffer from stiff neck a doctor observes that a patient has stiff neck what is the probability that the patient has meningitis. So, A we have this situation here a patient have come with stiff neck and we would like to see what the probability is of the patient having meningitis. So, the answer to that is obtained by the probabilistic technique, we need to find out probability of meningitis given s that is stiff neck. So, probability of meningitis given the stiff neck m and s are symbols for the random variables standing for having meningitis having stiff neck respectively.

(Refer Slide Time: 15:34)

Apply Bayes Rule (why?)

$$P(m|s) = \frac{P(m) \cdot P(s|m)}{P(s)}$$

$P(m)$ = prior probability of meningitis
 $P(s|m)$ = likelihood of stiff neck given meningitis
 $P(s)$ = Probability of stiff neck



So, we will apply Bayes rule, but we will keep in mind as to why we are applying this Bayes rule could we not have a worked directly with something which computes $P(m)$ given s directly we would like to apply Bayes rule. Now, $P(m)$ given s is nothing but $P(m)$ the prior probability of meningitis into probability of stiff neck given meningitis divided by probability of stiff neck.


So, first answer to our question as to why we should apply Bayes rule is this fact that as the problem has specified as the problem has specified we are given the probability of meningitis. We also see in the problem the probability of stiff neck meningitis and the probability of stiff neck all these are coming directly from the problem description. So, $P(m)$ prior probability of meningitis $P(s|m)$ likelihood of stiff neck given meningitis $P(s)$ probability of stiff neck, so all these values are known to us.

(Refer Slide Time: 16:43)

Probabilities

$$P(m) = \frac{1}{50000} \quad \text{Prior}$$
$$P(s) = \frac{1}{20}$$
$$P(s|m) = 0.5 \quad \text{Likelihood}$$
$$P(m|s) = \frac{P(m)P(s|m)}{P(s)} = \frac{\frac{1}{50000} * 0.5}{\frac{1}{20}} = \frac{1}{5000} \quad \text{posterior}$$

$P(m|s) \ll P(\sim m|s)$ Hence meningitis is not likely



So, $P(m)$ is 1 by 50,000 which is the prior probability we said that 1 in 20 have stiff neck. Therefore, prior probability of stiff neck is 1 in 21 by 20 and the likelihood of probability of stiff neck given meningitis is 50 percent which is 0.5. So, from this we can very easily compute the value of $P(m|s)$ $P(m|s)$ is equal to $P(m)$ into $P(s|m)$ divided by $P(s)$. So, 1 by 50,000 into 0.5 divided by 1 by 20 which comes out to be equal to 1 by 5,000, so there is 1 in 5,000 chances that the person has meningitis. So, from this it is easy to deduce that probability of meningitis given stiff neck is much less than probability of not meningitis given stiff neck.

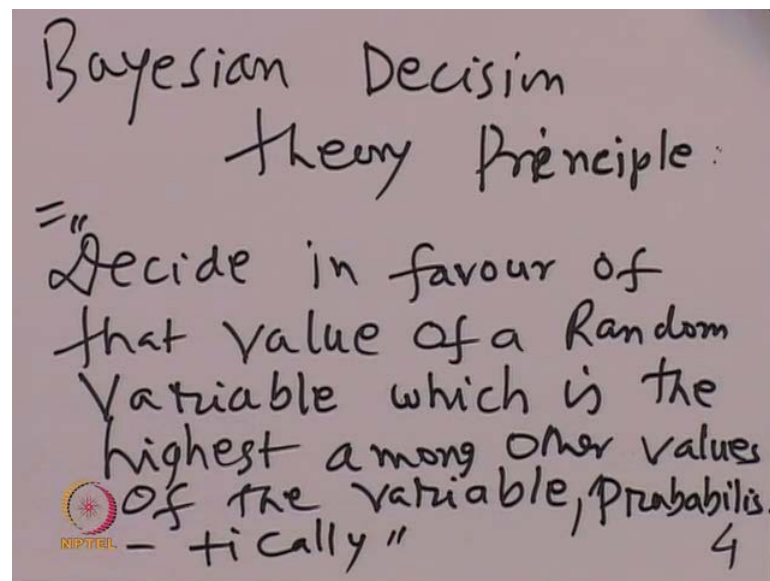
So, if we compare this to probabilities probability of meningitis given stiff neck and probability of not meningitis. Given stiff neck, we find that the probability of meningitis given stiff neck is much smaller 1 by 5,000 and hence meningitis is not likely. So, we should very carefully understand this problem there are many important things which I have shown by this example first thing that is shown is that look at this last line of the discussion $P(m|s) \ll P(\sim m|s)$.

So, this is helping us to make a decision whether the patient has meningitis or not, so this point is to be appreciated here our decision is purely a quantitative decision, the decision is obtained from comparing probability of $m|s$ with probability of not $m|s$. Given s , so our decision was a two way decision we wanted to know a patient arriving with stiff neck, thus the patient have meningitis or not. So, there are two ways there is a two way

decision making process in meningitis and the conditioning variable that is given is stiff neck. Now, we find the probability of meningitis given stiff neck from all the given data as specified in the problem and we find that this probability is less than the probability of not meningitis given stiff neck.

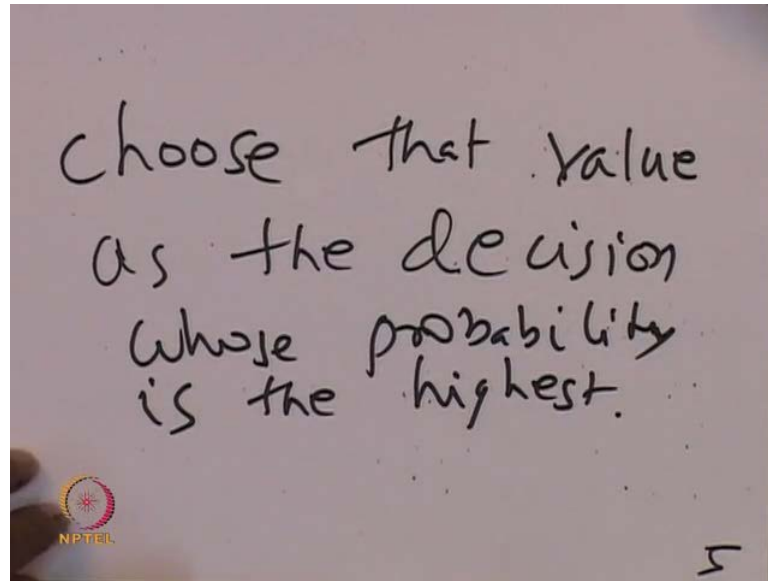
So, since this probability is less, therefore the other probability influences our decision we say that it is very unlikely the patient does not have meningitis with high probability. Notice the way the decision is being made the decision is made based on comparing two probability values and when the probability values have been computed. We are in a position where we can say probabilistically speaking or there is very good evidence that the patient does not have meningitis. So, this is the heart of Bayesian decision theory Bayesian decision theory says compare two probabilities. Compare a set of probabilities of the values of a random variable and pick up that particular a random variable as the decision or pick up that particular value as the decision whose probability is the highest.

(Refer Slide Time: 20:34)



So, we write down this Bayesian decision theory principle, so Bayesian decision theory principle says that decide in favor of that value of a random variable which is the highest among other values of the variable. So, this is the principle there is random variable it has many values a particular value decide in favor of that value of random variable. This is the highest among other values of the variable probabilistically that we are that here probabilistically choose that value as the decision whose probability is the highest.

(Refer Slide Time: 21:52)



So, let me write that sentence also very neat statement choose that value as the decision whose probability is the highest this is the essence of Bayesian decision theory, choose that value as the decision whose probability is the highest. So, we go back to the slide and see that we have made a decision with respect to the patient having meningitis or not. What we have made use of is the prior probability of meningitis which is very small 1 by 50,000, the probability of stiff neck have when meningitis is present. This is the likelihood probability 0.5 and probability of stiff neck stiff neck which is 1 in 20 and all this gives us this value 1 by 5,000 and this the probability balance in favor of lower meningitis.

(Refer Slide Time: 22:59)


Some Issues

- $p(m/s)$ could have been found as

$$\frac{\#(m \cap s)}{\#s}$$

Questions:

- Which is more reliable to compute, $p(s/m)$ or $p(m/s)$?
- Which evidence is more sparse, $p(s/m)$ or $p(m/s)$?
- Test of significance: The counts are always on a sample of population. Which probability count has sufficient statistics?



Now, we take up some issues, we could have found out prior probability of meningitis given stiff neck as a ratio of frequencies. Essentially, what we do is that we find out the number of people having stiff neck and then we find out of those patients, I mean stiff neck how many have meningitis. So, joint occurrences of meningitis in stiff neck we compute that number or observe that number divide it by the number of people having stiff neck. So, number of people having stiff neck both meningitis and stiff neck divided by number of people having stiff neck, so this whole ratio gives as the probability of meningitis given stiff neck.

So, now we ask which is more reliable to compute probability of stiff neck given meningitis or probability of meningitis given stiff neck which evidence is sparser. The probability of stiff neck given meningitis or probability of meningitis given stiff neck, so this means what evidence is larger in number. The larger the number, the more confidence we have in our observation and finally the test of significance the counts always on a sample of population which probability count has sufficient statistics.

The point we made is that when we compute these probabilities we compute these probabilities based on a population a sample of the population, so what is our confidence in saying that this probability holds over the whole general population that is the point. So, this is the test of significance when we discuss the questions there is a reason why we have been working with Bayesian theorem and we are what with the probability of P s

given m and not P_m given s . So, P_s given m is a smaller sized population we have to observe a smaller number of cases because meningitis itself is a rare about one in fifty thousand get this disease.

So, you take we take this small population and from this population find out how many have stiff neck. So, this is not the same as computing P_m given s in for P_m given s we have to take that population which as stiff neck and among this the find those which have meningitis. So, this is a much larger sample, we have to take a much larger group of people because stiff neck is much more common than a meningitis itself, but given that a person has meningitis that makes our sample a way definite set.

In that we observe the occurrence of stiff neck and therefore, these probabilities distinctively seems to be easy to compute and one can have more confidence in it. If you talk to medical professionals and doctors they would place their bets more on computing P_s given m rather than P_n given s and apply the Bayes theorem. We also have the advantage of working with the P_m value namely the prior probability of meningitis.

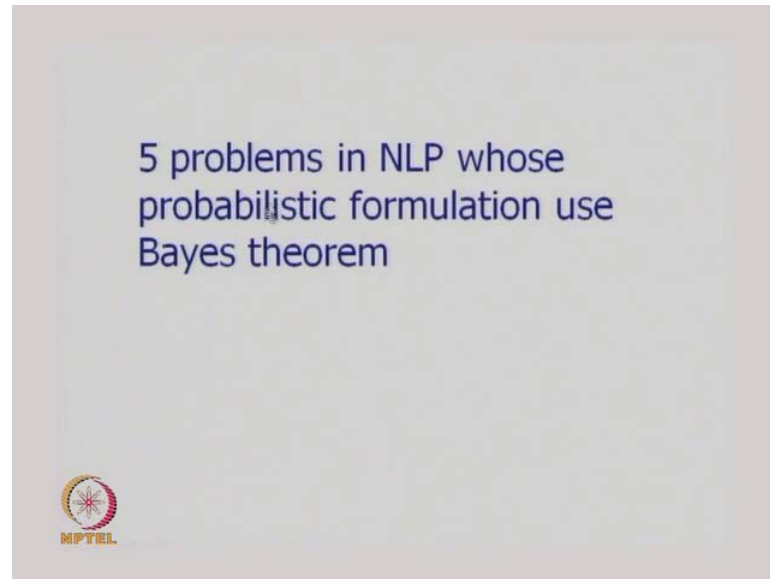
So, I spent so much time discussing the example mainly to bring home the following point the point is this when we apply statistical technique? Now, as a challenge model argmax computation etcetera we have to be very sure about which probability direction we are more comfortable or more confident of working with. So, is P_a given b or P_b given which has better a confidence in their values and which is easier to what which is easier to compute now just a point about the issue of significance we were talking about that.

Now, since P_m given x P_m given a s is a is not such an reliable parameter when computed directly because we have to directly observe a very large number people lots of people of stiff neck and that is also an amount of subjectivity involved in stiff neck. So, a stiff neck is sometimes is a feeling rather than a measurable medical condition where as there is meningitis e has very definite tests.

Through that, one can establish that the person has meningitis sop m given s a you necessarily has to be computed from a sample and can we take this probability to the whole general population. We can say that this probability holds over the whole general population our confidence in that is not likely to be as high as our confidence is probability of stiff neck given meningitis.

This probability is more likely to hold over the whole population of course here we are making some statements which are actually to be established must more rigorously to what is called the test of say significance. In the course we would like to discuss the standard techniques for test of significance, so this is the point a we have to see which direction of probability is convenient for us.

(Refer Slide Time: 29:35)



Now, we proceed and go to five problems in natural language processing whose probabilistic formulation makes use of Bayesian theorem. As we discuss the problems, we would also like to see if Bayesian theorem application of Bayesian theorem is a good idea in this case should we apply Bayesian theorem.

(Refer Slide Time: 30:02)



The problems

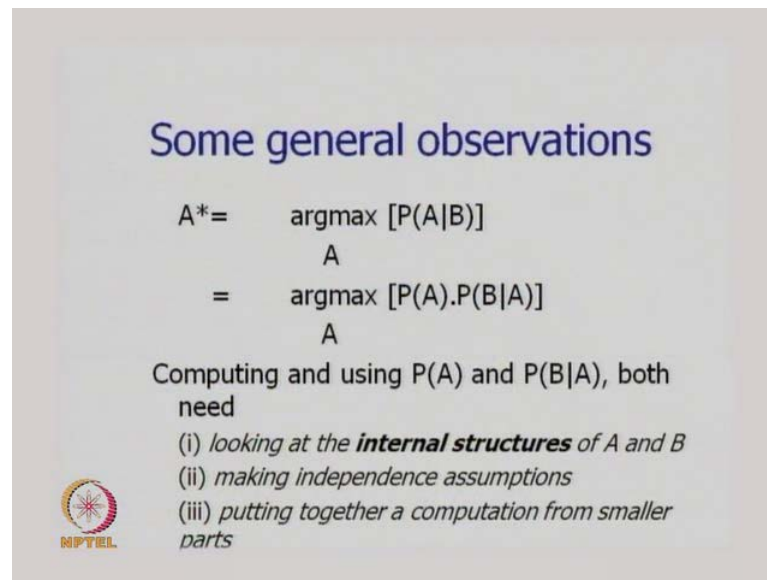
- Part of Speech Tagging: *discussed in detail in subsequent classes*
- Statistical Spell Checking
- Automatic Speech Recognition
- Probabilistic Parsing
- Statistical Machine Translation

NPTEL

So, let us discuss these problems the problems which are taken up are part of speech tagging. We will discuss this in much more detail in subsequent classes statistical spell checking, automatic speech recognition, probabilistic parsing, and statistical machine translation. On the phase of it might seem these problems are very desperate we did not see much similarity between two problems. What is the similarity between speech part of speech tagging and a spell checking or part of speech tagging automatic speech recognition or even probabilistic parsing and statistical spell checking? Between them how is statistical machine translation similar to let us say statistical spell checking.

So, this question arises what will see is that all this problems are actually some form of sequence labeling problem. In all cases we have a linear sequence of items which need to be given labels and when we give labels we would like to do the statistically. This statistical process is essentially an application of Bayesian theorem as we will see and in all cases we produce the best possible tag sequence or best possible label sequence given the linear sequence of items.

(Refer Slide Time: 31:37)




Some general observations

$$A^* = \underset{A}{\operatorname{argmax}} [P(A|B)]$$
$$= \underset{A}{\operatorname{argmax}} [P(A) \cdot P(B|A)]$$

Computing and using $P(A)$ and $P(B|A)$, both need

- (i) looking at the **internal structures** of A and B
- (ii) making independence assumptions
- (iii) putting together a computation from smaller parts



Let us see tag sequence or best possible label sequence given the linear sequence of tokens or items. Let us see what is the problem of this argmax based computation there are some inside pool general observations so we are interested in A^* where A^* is obtained by an argmax computation on $P(A|B)$ over all possible A 's. So, we find out the best possible A^* given the argmax through the argmax computation of $P(A|B)$ given B . We may apply Bayes theorem here which makes it $P(A|B)$ into $P(A) \cdot P(B|A)$ one might wonder what happens to the denominator because this whole thing is divided by $P(B)$.

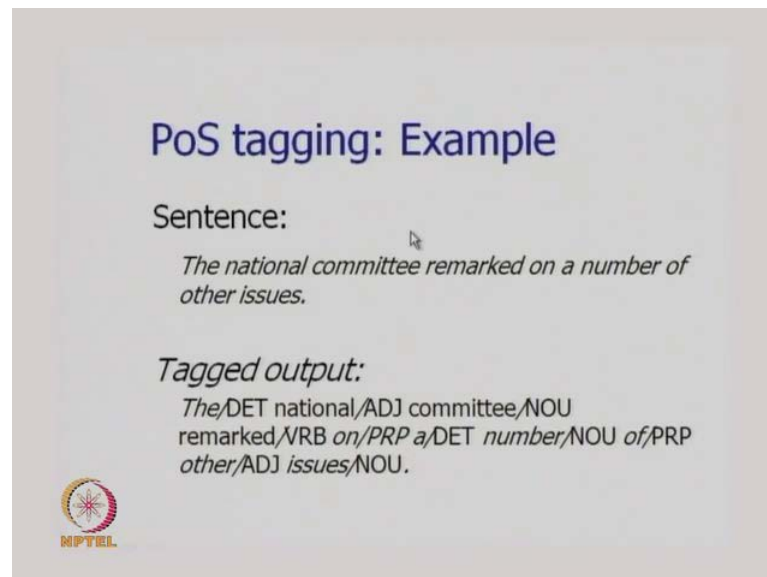
Notice that the argmax is computed over different values of A the value of $P(B)$ does not influence these argmax decisions in any way. Therefore, $P(B)$ can be dropped from the computation so without A^* , we can find it out through the product of $P(A)$ and $P(B|A)$ given A and then argmax computation on this quantity.

So, computing and using $P(A)$ and $P(B|A)$ both need looking at the internal structures of A and B because these are long sequences, and they have to be computed by dividing them into small parts. So, this means we need to look at the internal structures of A and B we have to make convenient and judicious independence assumptions, and we finally have to put together a computation from smaller parts.

So, let me just repeat these three points which is at the heart of statistical processing argmax based computation we have to compute the A^* we have to look at the internal

structure of A and B we have to make independent assumptions. Finally, we have to put together A computation forms smaller parts, so let us remember these three very fundamental points about argmax computation.


(Refer Slide Time: 34:10)



PoS tagging: Example

Sentence:
The national committee remarked on a number of other issues.

Tagged output:
*The,DET national,ADJ committee,NOU
remarked,VRB on,PRP a,DET number,NOU of,PRP
other,ADJ issues,NOU.*



We move on to PoS tagging the task of PoS tagging is exemplified by a sentence, here the national committee remarked on a number of other issues. The national committee remarked on a number of other issues when we tag the sentence we obtain the following sequence the determiner. So, it is given the level DET national committee here national is a qualifier for the noun committee and this is an adjective. Therefore, national is given the level ADEJ adjective committee is a noun, so it is given the level n o u. This forward slash followed by three characters is the label which is given on the entity. In the sentence remarked is nothing but a word, so slash VRV on is a preposition therefore, PRPA is again a determiner.

So, DET number is a noun on a number of issue, so number is noun NOU of is again a preposition PRP other is an adjective. So, ADJ and issues is an noun, so see how the whole sentence the national committee remarked on a number of the issues has been labeled with PoS tags like DED, ADJ etcetera, so let me point. You the fact that a theme tags have been found through the properties of this words and their function in the sentence and the relationship with respect to each other. So, committee is definite noun national is an adjective, but there are cases where their adjectives function as noun.

Therefore, disambiguating between adjective and noun deciding whether this particular entity should be an adjective or noun requires this ambiguities. So, all these labels which I have placed here they are nothing but the part of speech tags and they need to be computed with accuracy.


(Refer Slide Time: 36:40)

POS Tagging

Best tag t^* ,

$$t^* = \arg \max_t P(t | w)$$

$$t^* = \prod_1^{N+1} P(t_i | t_{i-1}, t_{i-2}) P(w_i | t_i)$$

 NPTEL

How do we form this problem as a sequence labeling problem through an argmax based computation? So, we say that we are interested in the best possible tagged sequence t star those are tags noun adjective etcetera tags. So, t star is the best possible tag sequence and that is found by an argmax computation on probability of t given w over all possible t argmax P t given w over all possible t .

So, this whole thing can become this whole thing can be a converted to a set of probabilities of this form and I am not discussing the independent assumptions etcetera here because. This is a subsequent discussion in a later class in later one or two classes so t star can be found as the product of P t_i given t_i minus 1 and t_i minus 2 into P w_i given t_i going from 1 to n plus 1. So, let us not worry about the dexterities here except to say that a this a this probabilities P w_i given t_i is nothing but probability of a wall w at a position i given that a tag at that position is t_i . Similarly, this is this first probability of t_i given that the tags at previous two positions are t_i minus 1 and t_i minus 2.

So, this probability is a conditional probability which is like a combination of three things so t_i t_i minus 1 t_i minus 2 a tag the previous tag and previous to previous tag.

So, this whole combination of three things a sequence of three things. It is known as the trigram, so we are computing the probability of a trigram because this whole thing $P(t_i \text{ given } t_{i-1} \text{ comma } t_{i-2})$ is nothing but probability of $t_i \text{ given } t_{i-1} \text{ } t_{i-2}$ divided by probability of $t_{i-1} \text{ } t_{i-2}$. So, a trigram probability divided by the bigram probability how we arrive at this formulation, let us hold it back for some time until later class.

Now, we see that this whole sequence labeling problem in the form of a tag sequence or a word sequence has been converted to an argmax based computation the question. We applying the the Bayes theorem here, because after Bayes theorem application these probability t given w actually becomes $P(t)$ which is nothing but the probability of the tag. Sequence prior probability of the tag sequence into probability of w given t probability of w can be ignored because the argmax is on t . So, we have applied Bayes theorem and a question we are asking is Bayes really necessary to be applied here.

(Refer Slide Time: 40:49)

$$t^* = \operatorname{argmax}_t P(t/w)$$

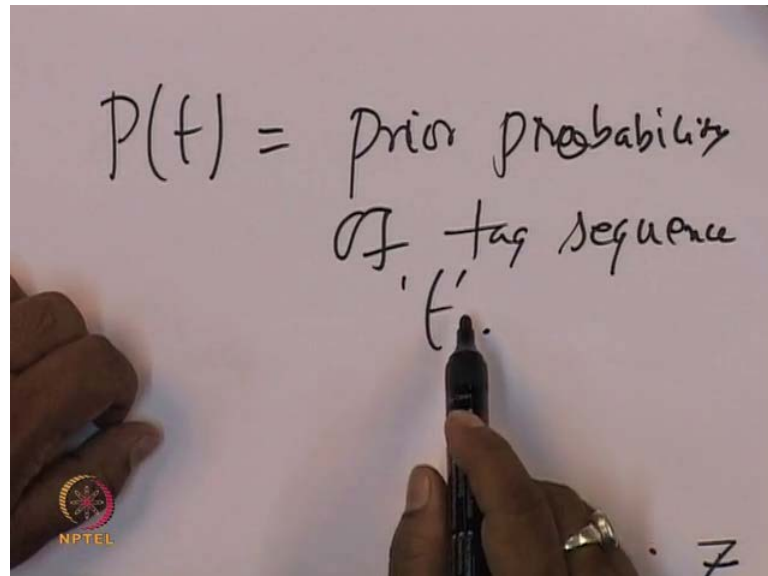
$$\operatorname{argmax}_t [P(t) \cdot P(w/t)]$$

$P(w)$ can be ignored.

So, we will do a bit of writing on this to understand this issue, so you can see the writing and see that t^* which is nothing that $\operatorname{argmax}_t P(y \text{ given } w)$ is converted to $\operatorname{argmax}_t P(t)$ into $P(w \text{ given } t)$ a $P(w)$ can be ignored. Now, the question is cant we work only with $P(t \text{ given } w)$ and a not apply Bayes theorem and work with $P(t)$ into $P(w \text{ given } t)$. The reason why we have applied Bayes theorem and we have converted the probability

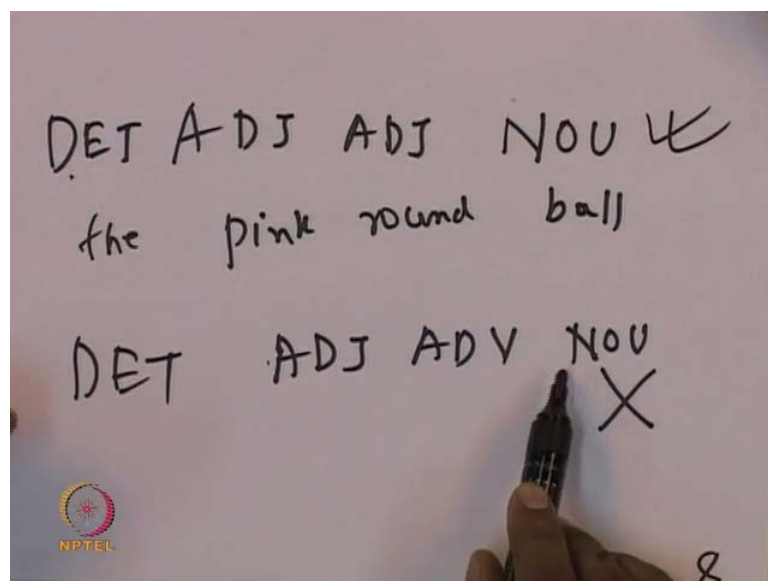
into a product of these two probabilities is that we get the valuable information from the prior probability $P(t)$, from the prior probability $P(t)$ it is as follows.

(Refer Slide Time: 42:06)



This $P(t)$ is equal to prior probability of tag sequence t , so probable is this tag sequence t . So, we are essentially trying to take advantage of more frequent tag sequences to take an example.

(Refer Slide Time: 42:44)



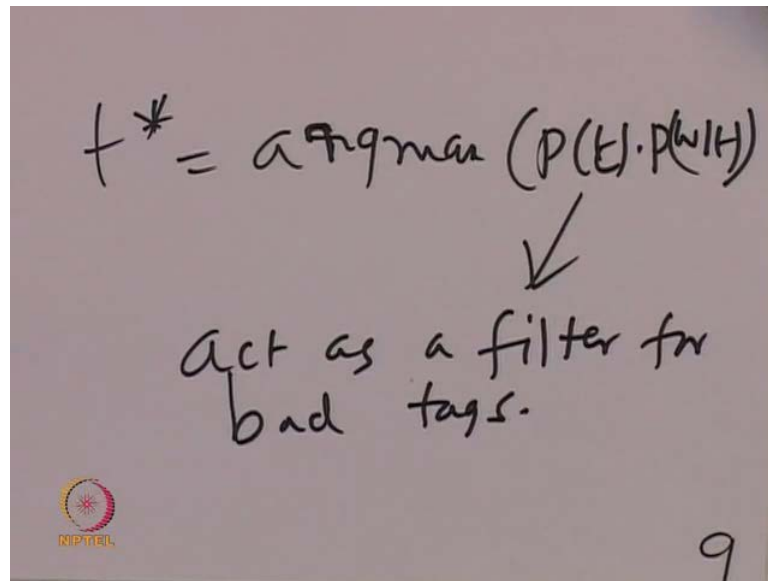
Suppose, we have ADJ, ADV and NOU sequence, so look at the sequence adjective noun sequence an example of that would be let me. Similarly, ADJ, ADJ NOUS an

example of that would be the pink round ball the pink round ball. So, this kind of sentence this kind of sequence of tags is this sequence of tags is been very common determiner extreme of adjectives. Then, a noun this is very common this is more common for example, and then a sequence like DET, ADJ, ADV and a noun. I think you will find it very difficult to get a tag sequence such as that you have a determiner the adjective then an adverb and then a noun.


So, this is very unlikely because the adverb typically comes before an adjective and that too in a very restricted way you can have for example, very good is unlikely. So, an adverb qualifies either a verb or an adjective, so in English since the qualifier typically comes modifier typically comes before the modified adverb for following an adjective is a very unlikely sequence.

Then, if a noun is coming after that that makes it all the more unlikely, so adverb coming before a verb is more likely adverb coming before a noun is much less than likely he quickly drove he quickly drove away. So, quickly drove or quickly drove away here is an example that is an example of adverb coming before a verb adverb coming before noun is not so common. It is still much less common to have an adverb flanked by an adjective and noun on to two sides I do not think you can find any example easily. So, this sequence is common this sequence is not so common, when we do the computation and produce the tag sequence by means of a computation by means of the argmax computation.

(Refer Slide Time: 45:51)

A handwritten slide on a light-colored background. At the top, the equation $t^* = \text{argmax} (P(t) \cdot P(w|t))$ is written in black ink. A downward-pointing arrow is drawn below the equation. Below the arrow, the text "act as a filter for bad tags." is written in black ink. In the bottom left corner, there is a small circular logo with a red and yellow sun-like symbol and the word "NPTEL" underneath. In the bottom right corner, the number "9" is written in black ink.
$$t^* = \text{argmax} (P(t) \cdot P(w|t))$$

act as a filter for bad tags.

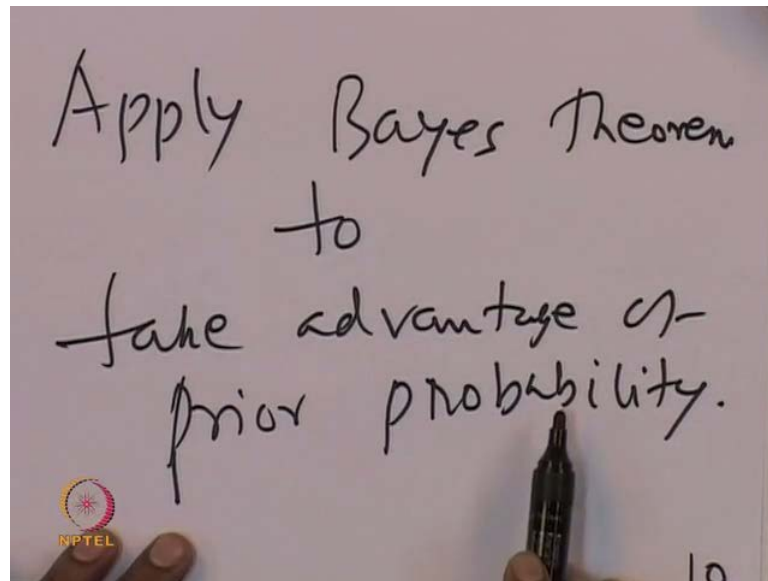


9

Now, what we find is that some sequences more likely than others and the $P(t)$ value which comes $\text{argmax} P(t) P(w|t)$ given t , the $P(t)$ well which comes that can act as a filter for bad tags. So, I think now the motivation for applying the Bayes theorem is quiet clear we would like to catch those tag sequences which are unlikely and that is made very explicit by means of a prior probability component.

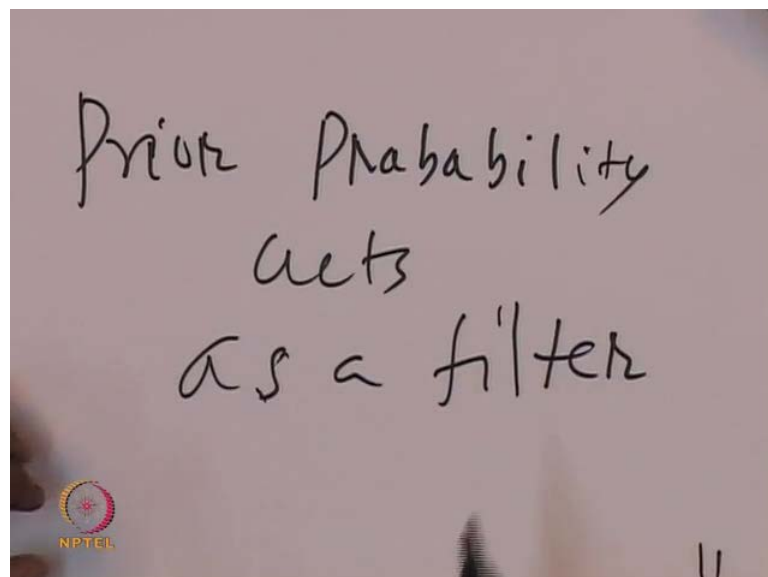
In the computation prior probability component in the form of $P(t)$ and this $P(t)$ is very low this $P(t)$ is very low for unlikely tag sequences and that will make the t^* for a unlikely tag sequences a not obtainable to the argmax computation. So, let us remember this example that we apply Bayes theorem to take advantage of prior probability, so let me write it down as an important principle in Bayesian theorem application.

(Refer Slide Time: 47:27)



Apply Bayes theorem to take the advantage of prior probability, so if we are applying Bayes theorem then the prior probability gets separated from the rest of the formulation. It becomes exquisite and one can make use of this as a filter, so let me write this important statement also.

(Refer Slide Time: 48:09)



Prior probability acts as a filter, this is such a common idea and such useful idea that it finds its appearance in a number of NLP statistical NLP situations. That prior probability being exquisitely available through Bayes theorem makes some sequences which are

unlikely high probable. That is the reason why Bayes theorem is used, in the next lecture we will see a more examples of these were Bayes theorem is used making exclusive the prior probability and making its use and explicit filter.