Natural Language Processing

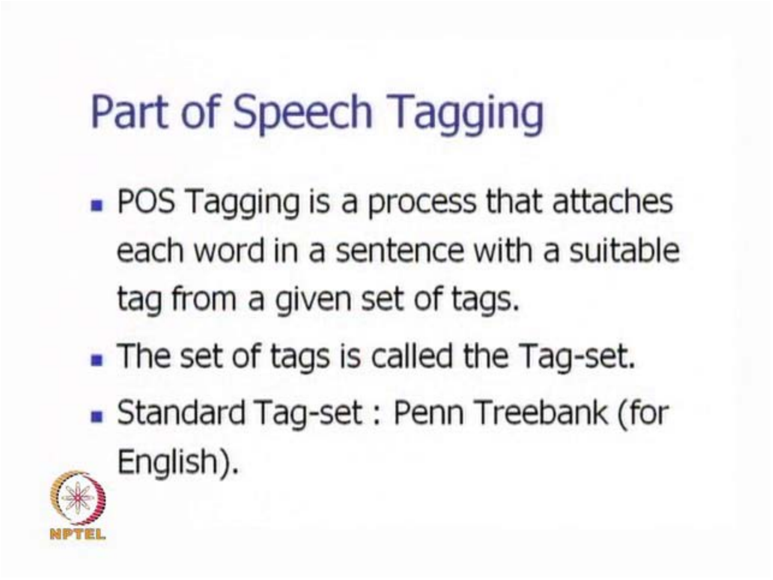
Prof. Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay

Lecture - 10
Part of Speech Tagging

Today, we are going to talk about Part of Speech Tagging, which is an extremely important topic, in statistical natural language processing or let us say a whole of natural language processing. And the fact that part of speech tagging is an important component of natural language processing is well recognized. But the computational task or part of speech tagging gained visibility and respectability, let us say in a recent times, in last 10 to 15 years, but it is very, very well understood that, a good part of speech tagger is a crucial component of any natural language processing task.

Before beginning any processing, any sophisticated processing on text, it is important to know the categories of the words that form the text. So, this is the discussion of today, we are going to describe the part of the speech tagging problem, the statistical approach to doing this, because it is well recognized again that part of speech tagging is vest done by machine learning methods from the annotated corpora. Human created rules for part of speech tagging do a good job no doubt, but they are little subject to human error and may miss out phenomena. But on the other hand, if you have data or text, which is already marked with part of speech tags. And we run a machine running algorithm on this, then it is possible to create a statistical system which does a very good job of part of speech tagging.

(Refer Slide Time: 02:14)

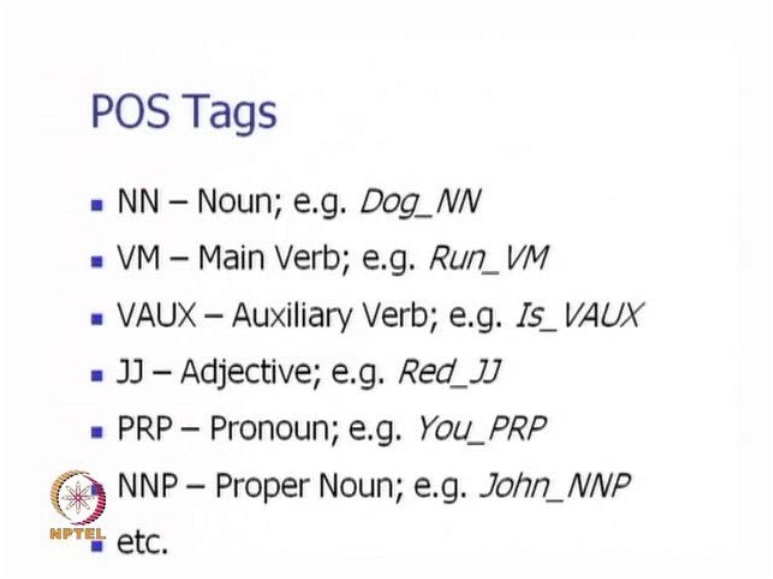


Part of Speech Tagging

- POS Tagging is a process that attaches each word in a sentence with a suitable tag from a given set of tags.
- The set of tags is called the Tag-set.
- Standard Tag-set : Penn Treebank (for English).

So, here is the definition of the problem, part of speech tagging is a process that attaches each word in a sentence with a suitable tag from a given set of tags. The set of tags is called the tag set, there are many standard tag sets Penn tree bank for English is probably the most visible and famous set of tags for categorizing the words of English language.

(Refer Slide Time: 02:47)



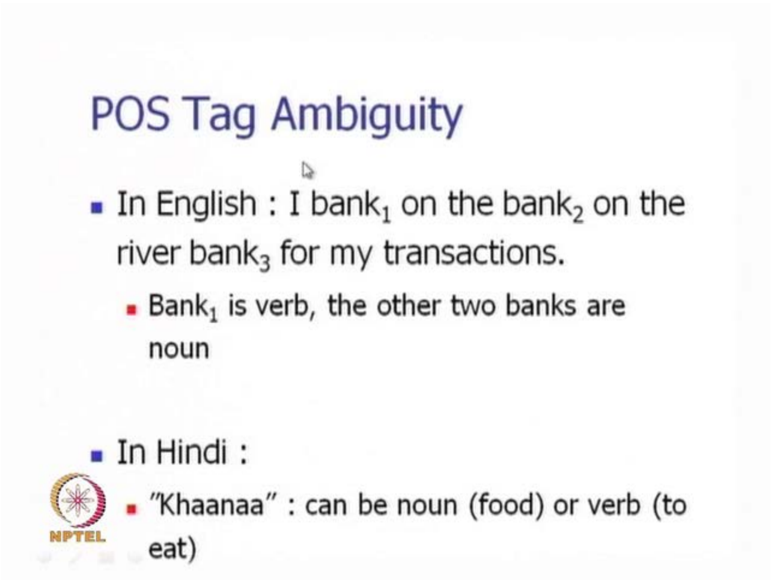
POS Tags

- NN – Noun; e.g. *Dog_NN*
- VM – Main Verb; e.g. *Run_VM*
- VAUX – Auxiliary Verb; e.g. *Is_VAUX*
- JJ – Adjective; e.g. *Red_JJ*
- PRP – Pronoun; e.g. *You_PRP*
- NNP – Proper Noun; e.g. *John_NNP*
- etc.

Now, here is some examples of part of speech tags, N N is noun for example, Dog is N N, we had discuss this in the last class, but let me repeat, this point once again just to clarify the notion of tags and how they are placed with the words. So, one of the options to place the category information is by placing an under score and then placing N N, but you could also conceive of this being and x m l file, where each word is place within x m l mark, which also specifies what the tag of the word is anyway those are representations. However, we find that, each word needs a categorization in the text, we do not leave out any word. Similarly V M, which is a very important tag, namely the main verb of the sentence for example, john runs here run is the word and it has the category V M main verb, which an underscore, V AUX is the auxiliary verb, so is am or these are auxiliary verbs, they also a tag in the text.

So, is underscore V AUX is the tag shown here, J J is the adjective tag and I remark that, adjective is sometimes pronounced to it d very similar to the J sound. So, Red is an adjective as in Red ball, so Red underscore J J would be the tag for the word Red P R P is also an important tag pronoun. So, you is a pronoun, so you underscore, P R P is the tagged word, N N P is the tag for proper noun john underscore N N P, shows the tagging for john, as in the sentence john runs, so N N P is the tag for john and so on. So, this way a piece of text is completely tagged, for every word in it.

(Refer Slide Time: 05:06)



POS Tag Ambiguity

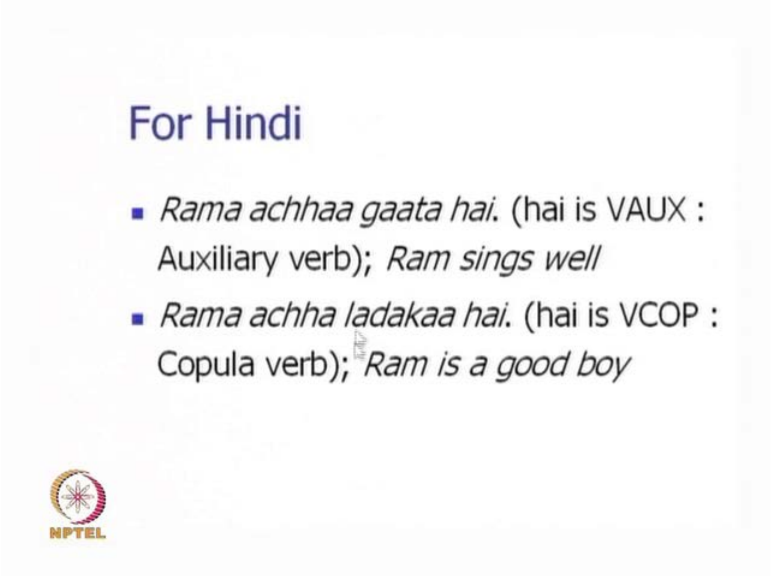
- In English : I bank$_1$ on the bank$_2$ on the river bank$_3$ for my transactions.
  - Bank$_1$ is verb, the other two banks are noun
- In Hindi :
  - "Khaanaa" : can be noun (food) or verb (to eat)

Proceeding further, now pos tag ambiguity is very common in text, I take her a sentence, which was mentioned in the last class to, in English I could have a sentence, I bank on the bank, on the river bank for my transactions. So, I bank on the bank on the river bank for my transactions, the lower suffixes indicate the occurrence of the word bank, so bank

1 here is a first occurrence of bank, bank 2 here is a second occurrence and bank 3 here is a third occurrence.

Bank 1 is verb, because I bank on the river bank for my transactions, yet this bank means depend, so I depend on the bank on the river bank for my transactions. So, bank 1 is verb, the other 2 banks or noun, I bank on the bank. So, this is the actual bank where financial transaction, take is taking place and this third bank is nothing but the river bank in Hindi, the word [FL] can be a noun or a verb. So, when it is a noun it means food when it is verb it means to eat, so [FL] again has pos tag ambiguity.

(Refer Slide Time: 06:35)



For Hindi

- *Rama achhaa gaata hai.* (hai is VAUX : Auxiliary verb); *Ram sings well*
- *Rama achha ladakaa hai.* (hai is VCOP : Copula verb); *Ram is a good boy*

Now, I take 2 examples here, which were mention towards end of the last class, but let us describe this in little more detail, just to bring out the ambiguity issue in POS tagging ram [FL]. So, [FL] is auxiliary verb, it is an auxiliary verb, so it has the tag V AUX. The meaning of the sentence is ram sings well, now ram [FL] here this [FL] is a qualified for the verb and therefore, this is an adverb. So, ram [FL] is an adverb ram [FL], so see that both the sentences look very similar except for this 2 words [FL] and [FL]. Here, we say that [FL] is not an auxiliary verb, because it is not a helping verb, an auxiliary verb requires a main verb, which it is helping, in the previous sentence [FL] was the main verb and [FL] is the auxiliary verb, it is the helping verb for [FL]. The auxiliary verb, carries the tense number and person information ram [FL] here indicates that, the person is third person single or number and the tense is present tense ram [fl] ram sings well present tense singular number third person all this information is carried on hai.

Now ram [FL] here, [FL] is not preceded by a main verb, we do not call it an auxiliary verb, it is called a copula verb and the symbol for this in parts of speech tagging is V COP. So, ram [FL] indicates the goodness of ram, ram is a good boy, now here [FL] qualifies [FL], which is a noun therefore, [FL] is an adjective, in the previous sentence [FL] was an adverb, because [FL] was qualifying the verb. So, the whole point of this discussion is that, this sentence are very similar, except for this towards [FL] and [FL].

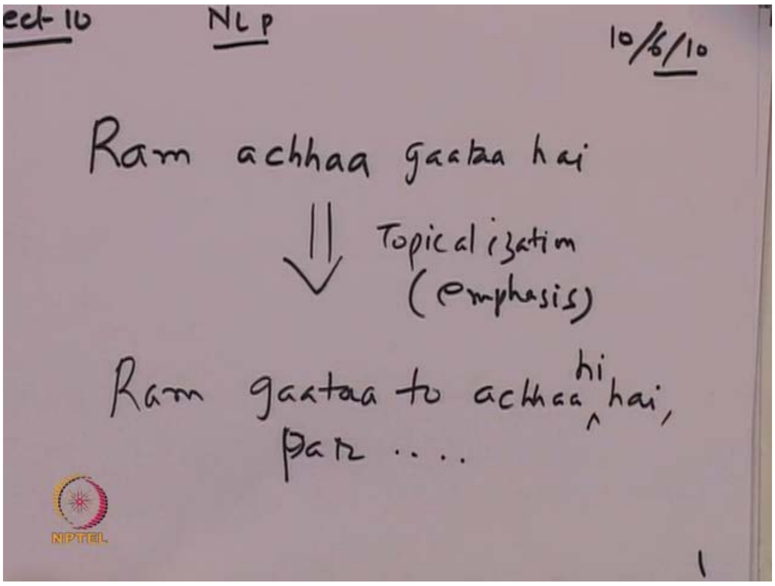
But, the word [FL] has different POS tags depending on what it qualifies in the first case, it is adverb, second case it is an adjective, similarly [FL] preceded by verb, which is helping the main verb is the auxiliary verb and [FL] here is the copula verb. Now, we can understand that, words can a multiple pos tags and it is important to dis ambiguity them, it is important to place the correct tag depending on the sentential context, in which it appears.

Now, if you look at this 2 sentences, you may be tempted to come up with this kind of rule that, when a word, which can be both adverb and adjective for example, [FL] here, if it followed by a verb [FL]. This then is an adverb, if it followed by a noun then it is an adjective. So, this is because the principle that is operating in your mind is that, [FL] is qualifying a verb therefore, it is adverb here [FL] is qualifying a noun therefore, it is an adjective here, but the problem comes when we insist that, the following word is the clue for disambiguation.

So, part of speech tagging of course, is the first level task, syntax semantics all this have not been done, we have just started processing the text. So, we cannot assume any syntactical clue or semantic clue, we cannot assume parsing is done or the semantic role has been obtained. So, we necessarily have to depend on the clues from nearby area, so nobody refuse that, nobody argues with that, it is defiantly the case that, we have to disambiguate vast on clues available in the near vicinity.

The most powerful clue comes from the suffixes the morphological features on the word itself, so in this case, we find [FL] is a word and the both cases the word forms the same, therefore there is no morphological clue, actually there should be an a here [FL]. So, ram [FL], there should be an a here and if we say that it is followed by a verb it is adverb followed by a noun, it is adjective. Now, the problem is that, it is possible to have some amount of text between [FL] and [FL] and [FL], so ram [FL] for example, ram [FL]. So, you have the particle he between [FL] and [FL] ram [FL], this is again possible and therefore, you can have some amount of text between [FL] and [FL] sometimes, what can happen is that, because of movement, because of topicalization focus etcetera, words can move ram [FL], you can say ram [FL] ram [FL] to [FL]. So, now, you can see here [FL] has gone after [FL] here to has gone after [FL], so let me write this to clarify.

(Refer Slide Time: 12:49)



ect-16 NLP 10/6/10

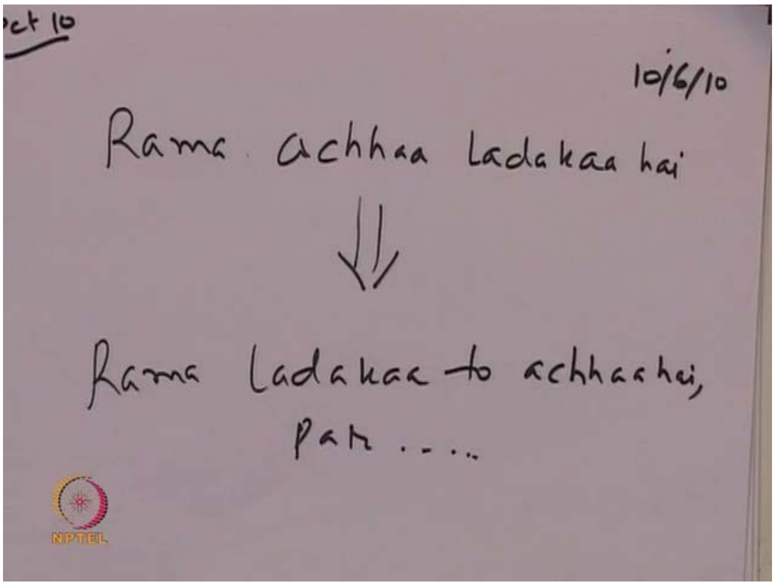
Ram achhaa gaataa hai

⇓ Topicalization (emphasis)

Ram gaataa to achhaa hi hai, par . . .

What I am saying ram [FL], because of topicalization, which means emphasis, one could have the situation ram [FL] to [FL] per and you can have some other piece of text coming after it. So, here also you can see that, [FL] is an adverb, it is still qualifying [FL], but if you if we say that, the rule is adverb should be followed by a verb, that rule will fail in this case ok. So, the rule will work in most cases from [FL] is adverb, this is fine, but in this case, because of the movement of the word, even though [FL] is adverb, this rule is not applicable, because [FL] is coming before [Fl], ram [FL] to [FL], we could also have ram [FL] to [fl] making it little more complex. Now can the same thing happen for adjective.

(Refer Slide Time: 14:12)



ct 10

10/6/10

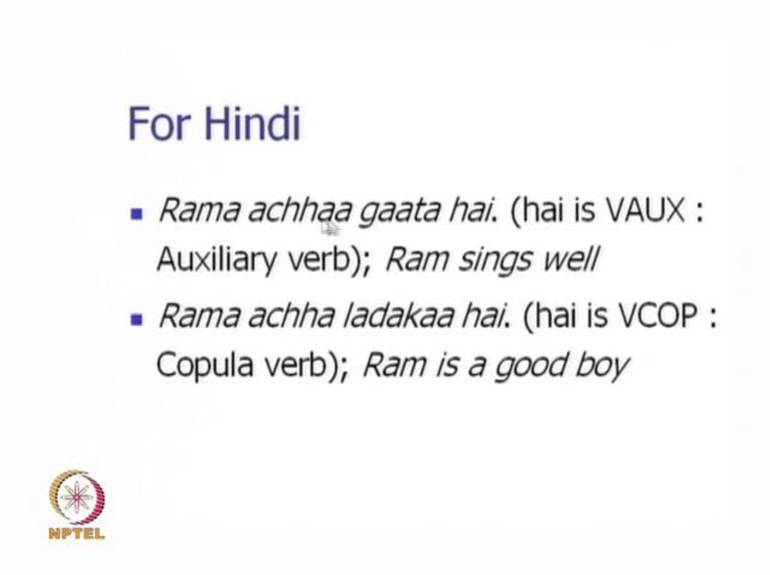
Rama achhaa ladakaa hai

⇓

Rama ladakaa to achhaa hai, par ....

So, we had the sentence ram [FL] again, because of word we meant, this rule of noun following an adjective will not work here, ram you can say ram [FL] par an another piece of text identical situation. So, [FL] has moved and because of this movement, the fact that, [FL] is qualifying a noun and therefore, it is an adjective that rule, we did not work. So, I suppose you have understood that this is the point.

(Refer Slide Time: 15:18)

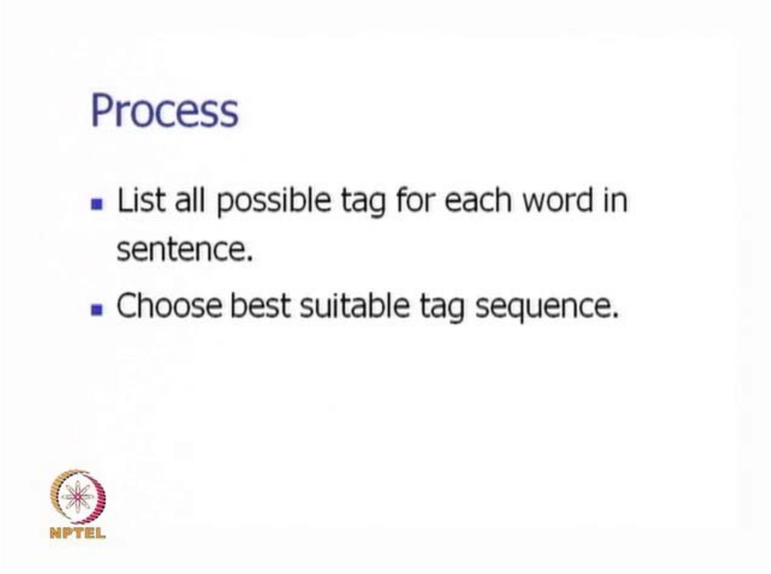


For Hindi

- *Rama achhaa gaata hai.* (hai is VAUX : Auxiliary verb); *Ram sings well*
- *Rama achha ladakaa hai.* (hai is VCOP : Copula verb); *Ram is a good boy*

We have to use simple rules for part of speech tagging, namely the word clues from the immediate vicinity of the word, but they are fallible rules, there vital rules, because there can be many natural language phenomena, which work against this rules.

(Refer Slide Time: 15:33)

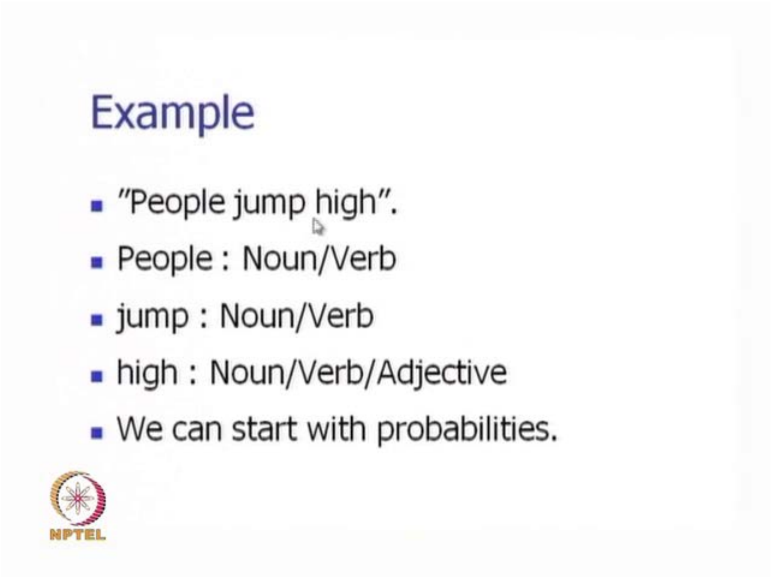


Process

- List all possible tag for each word in sentence.
- Choose best suitable tag sequence.

Proceeding further the process of part of speech tagging, is list all possible tag for each word in the sentence, choose the best suitable tag sequence. So, this is the process for each word will list all possible tags, for each word in the sentence and we choose the best suitable tag sequence.

(Refer Slide Time: 15:56)

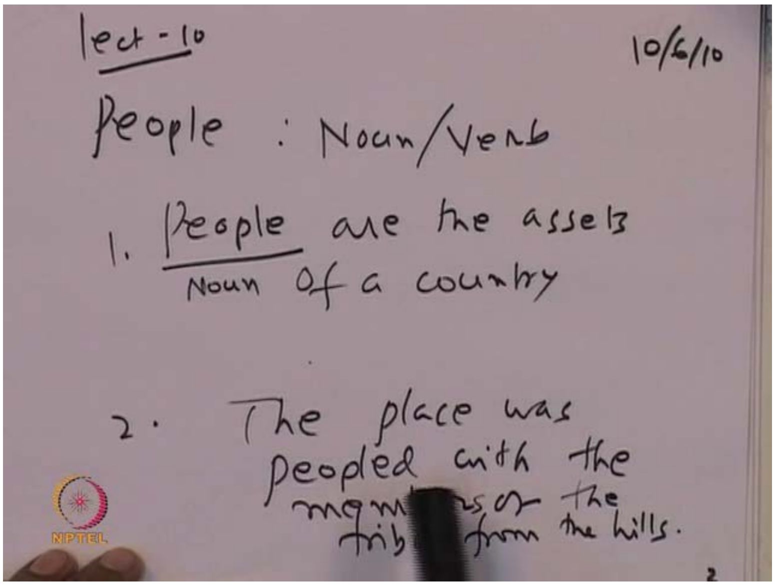


Example

- "People jump high".
- People : Noun/Verb
- jump : Noun/Verb
- high : Noun/Verb/Adjective
- We can start with probabilities.

Here is an example to illustrate, this point people jump high, it is a fictitious sentence, let us not worry about the meaning of this sentence as to what it could possibly indicate may be it is a sentence from a this course in a context. So, people jump high people can be both noun and verb, jump can be noun and verb high can be noun verb adjective. So, if you are not convinced, we can take examples to see, how people jump in high can have this multiple tags, this I believe needs a bit of explanation.

(Refer Slide Time: 16:39)



lect-10

10/6/10

People : Noun/Verb

1. People are the assets of a country

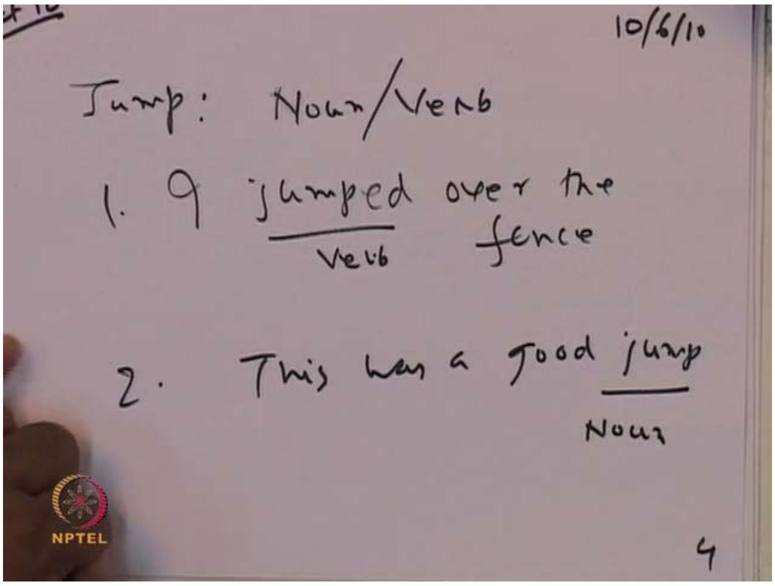
Noun

2. The place was peopled with the members of the tribes from the hills.

2

So, we take people can be noun or verb how can it be noun, that is quiet simple people are the assets of a country. So, here it is noun people can be verb the place was peopled with the members of the tribes from the hills, I hope you can read the sentence the place was peopled with the member with the members of the tribes from the hills. So, let us say a place has been built and we populate, the place with members of the tribes from the hills, so here people means populate it, so thus people can be both noun and verb let us see the other words in this sentence jump.

(Refer Slide Time: 18:09)



10/6/10

Jump: Noun/Verb

1. I jumped over the fence

Verb

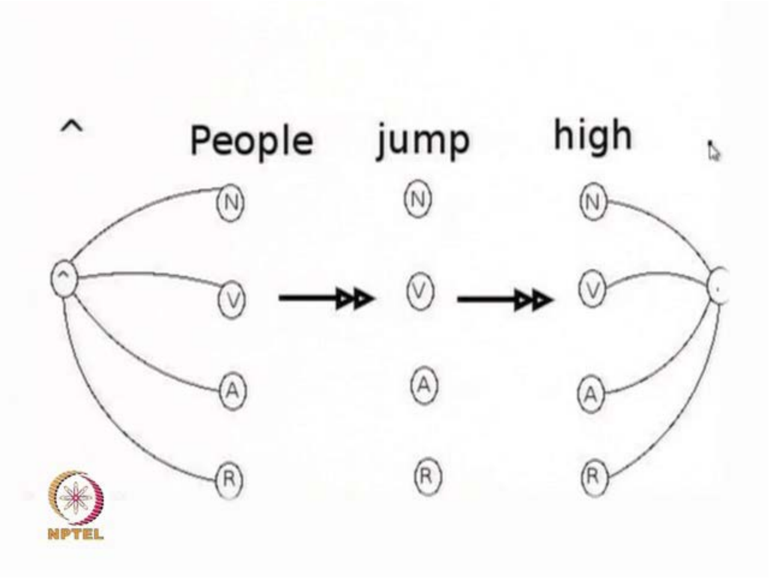
2. This was a good jump

Noun

4

Can be both noun and verb, this is relatively easier, the fact that jump is verb is well known, I jumped over the fence, so here this is a verb, this was a good jump here, it is a noun, so such sentence are quiet common. We can take the word high finally, high can be noun verb and adjective, I believe the verb parts of speech tag is very rare, we live out verb, but high can be noun and adjective, which is quiet common, so for example, high hills. So, here this is an adjective and after the win, he was on a high here this is noun, so after the win, he was on a high this is noun. So, thus we find that, what is a multiple parts of speech and it is quiet common to have words in a sentence with multiple part of speech and it is necessary to disambiguate them. So, we proceed further.

(Refer Slide Time: 19:42)



And now suppose, we have the task of part of speech tagging in front of us, we would like to pos tag, this words in the sentence people jump high. Now, a very useful convention is to have the sentence beginner, which is typically the hat symbol and the sentence finisher, the sentence ended, which is the dot symbol, this is the full stop. So, these 2 d limiters are the sentence for the sentence, the part of speech tagging begins from the hat ends on dot, for a particular sentence.

So, this shows here a very simple picture, we place the tags along site or on top of every word in the sentence. So, the people, we saw could be noun and verb just for a simple a scheme, just for this discussing a simple scheme. We place all the tags or all the words, so this is a simple minded scheme, but it will work, because the tags, which are completely improbable will not be taken up. How one could place tags, which are applicable is a separate discussion, which we can go in to later.

So, we have this 3 words people jump high or people, we have N V A and R, which are noun verb adjective and adverb tags, so all the words have this 4 tags, so assume in our discussion, we are concerned only with 4 tags noun verb adjective and adverb. So, we start with the hat symbol, which is the beginner of the sentence, from here, we can transit to N or V or A or R from here, we can transit to any of this 4 tags and this transition is from each state. So, each tag is a state here, so from this state 4 transitions are possible from, we state 4 again A again 4, R again 4.

So, there would be 16 arcs going from the this set of states, this first column of states to the second column of states, similarly from the second column to the third column, we can have 16 transitions, because from each state, we can make 4 transitions. And finally, we have this 4 transitions going into the finishing state, which is dot. So, the an important point here to notice that, the sentence beginner has this prevail tag of hat and the sentence ended or the full stop has this prevail tag of dot. So, this whole graph, which is nothing but a graph is to be traversed for best possible path from hat to dot.

So, now, let us discuss a an important point, what we have done, so for is that, we have taken the words of a sentence and we have erected columns of fast tags on them. It is a separate matter of discussion as to how, we can selectively place, only those tags, which are applicable, this is not very important right now. For understanding, what is going on as a process, it is sufficient for us to see that on top of words, we have tag sequences columns of tags, at each tag should be looked up on as a state. The first starting state is the hat state, which begins the pos tagging processes and the last finishing state is the dot state, which is the full stop. And this finishes the tagging process on the way starting from hat to the dot symbol, we are interested in finding the best possible path from hat to dot traversing the states, from each column, we choose only one state.

So, if there are n words in a sentence, there are 2 delivators hat n dot, so if there are n words then there are states of the tags and finally, when we find the best possible tag sequence, we would have described a path of length n plus 2 in the sense that, there are n plus 2 nodes in this whole path. And this path gives me the best possible tag sequence we choose one state, one single state from each column of states. So, we can see now that, the whole pos tagging process has been reduce to a graph traversal task, starting from the hat state to the goal state, we find the best possible path and this path chooses only one state, from the column of states on each word. So, this makes a formulation quiet clear, we are now ready to look at the techniques of fast tagging.

(Refer Slide Time: 25:14)

Challenge of POS tagging

Example from Indian Language

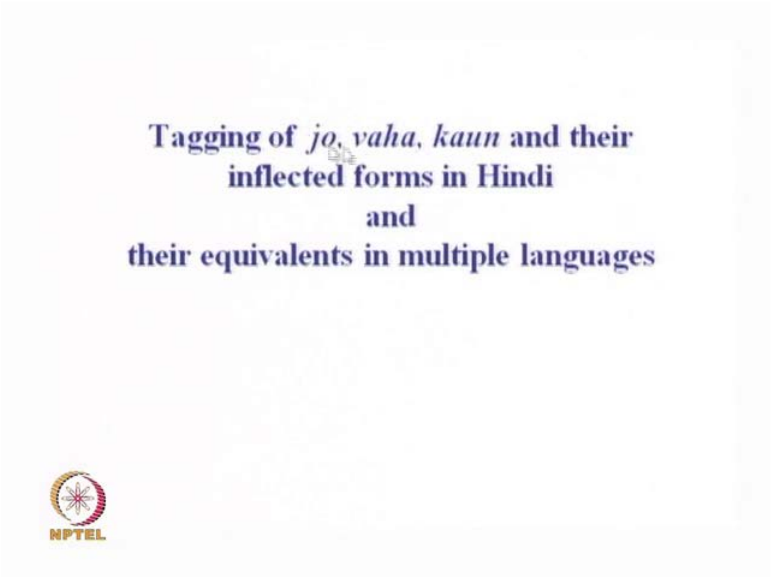
Before that, we would like to understand why pos tagging could be channeling and we will be discussing mainly with the English examples, but in this part let us take some examples from Indian languages and this would show, why part of speech tagging is a real challenge, it is not a trivial task. At this part let me just remind you once again this fact, we spend just 2 minutes to repeat a point, part of speech tagging is crucial for any L L P task, you have to begin natural language processing of text with part of speech tagging.

So, emphasis is on the word begin, we start natural language processing with part of speech tagging ok. So, before part of speech tagging, what is available possibly is morphological analyses information, so if a language has morphology analyzer, it would take the words of the language and stripe of the suffix get the morphological features of the words for example, the word [FL] here, the suffix is a the root what is [FL] the number is plural form.

So, this is the plural form of [FL], it could also with the oblique form, so let us not go into those details, a point is that, the words have been presets for separating the suffix from the root word and the word features are available. So, when you begin part of speech tagging the only information that is available is the information at the word level morphological features and the suffixes. So, we cannot assume any syntactic information is available, we also cannot assume that semantic roles are available.

Though the semantic properties of the words may be available, but that may require since this ambiguition, it may require words in this ambiguition, but words into this ambiguition is later task, which has to be done after part of speech tagging. Therefore, it is necessary the case, that part of speech tagging has to be done with tremendous amount of constraint, you can place part of speech tag on a word only from limited amount of context in the vicinity of the word. So, the word itself and may be some 2 or 3 words, before and after it that is all, this is the only information available and using that, your force to do that is ambiguition. So, this produces challenges and let us looks at some of these challenges, which are very interesting.

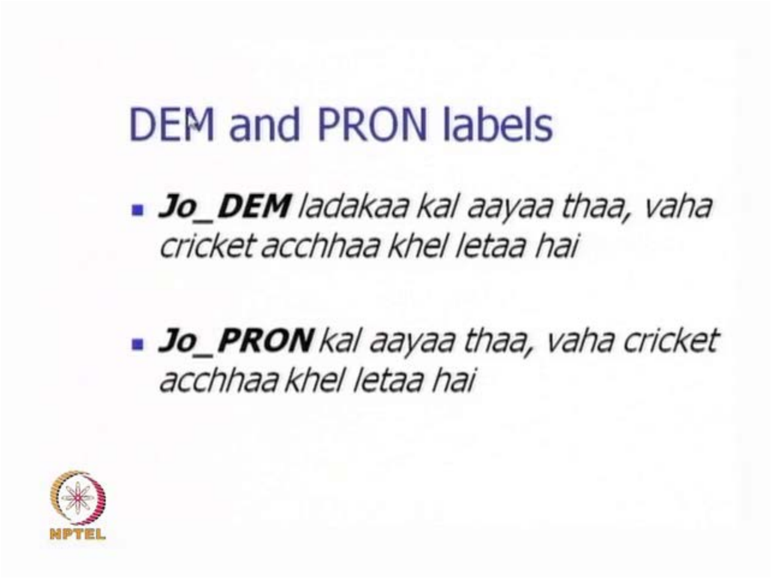
(Refer Slide Time: 27:55)



Tagging of *jo, vaha, kaun* and their inflected forms in Hindi
and
their equivalents in multiple languages

We take up the phenomenon of [FL] and their inflected forms in Hindi and their equivalents in multiple languages, we will mainly discuss Hindi at some example from Bengali and Sanskrit. So, the problem is to place tags on [FL] and [FL] on the text and their forms.

(Refer Slide Time: 28:25)



DEM and PRON labels

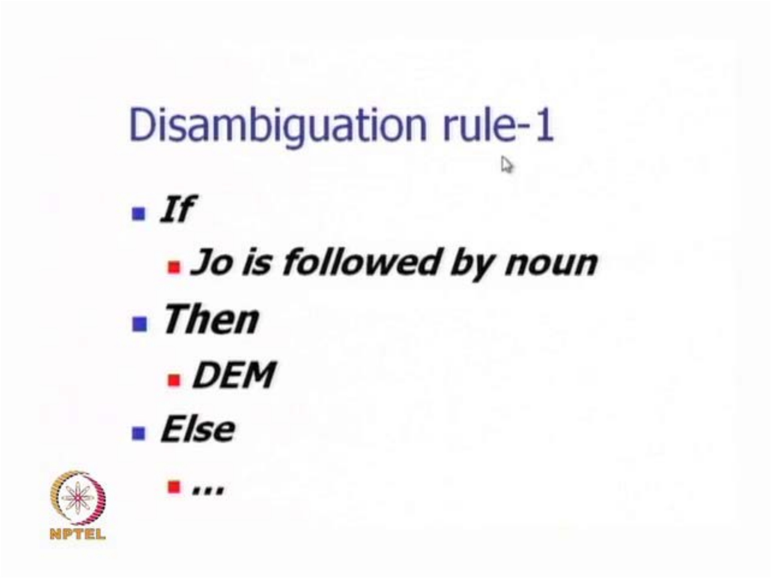
- ***Jo_DEM*** *ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*
- ***Jo_PRON*** *kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

Now typically, the level that is given are DEM and PRON, please understand, what we mean by this DEM means demonstrative and PRON means pronoun, so PRON is pronoun DEM is demonstrative. Let us look at this sentence [FL] cricket [FL], this is the sentence [FL] cricket [FL], the boy who came yesterday plays cricket well, this is the meaning of the sentence in English. Look at the word Jo here, which is in capital and it has been given the tag of DEM Jo underscore DEM, DEM is the tag on Jo, which indicates demonstrative.

So, here the word Jo has a demonstrative function Jo [FL] cricket [FL], we a specify particular boy [FL], so that is why, it is demonstrative. Take the next sentence [FL] cricket [FL], so almost same sentence expect that [FL] is dropped [FL] is dropped. So, Jo has to find, it is what is called referent, what does Jo refer to, it refers to something, which is not present in the sentence, however, this Jo can be matched with [FL] Jo [FL]. So, Jo [FL] who the person, who came yesterday plays cricket well, this is the meaning.

Now, this Jo not a demonstrative, its pronoun because it is reference is somewhere else and this Jo ending visiting has what is called demonstrative role, it indicates a particular boy Jo [FL] and this Jo has an unspecified noun another and it is a pronoun, so this Jo is a pronoun. Now it is clear that here, we are faced with a disambiguation situation, because Jo can have both done and from and we need to find out, which level will be applicable in the particular context.

(Refer Slide Time: 30:56)

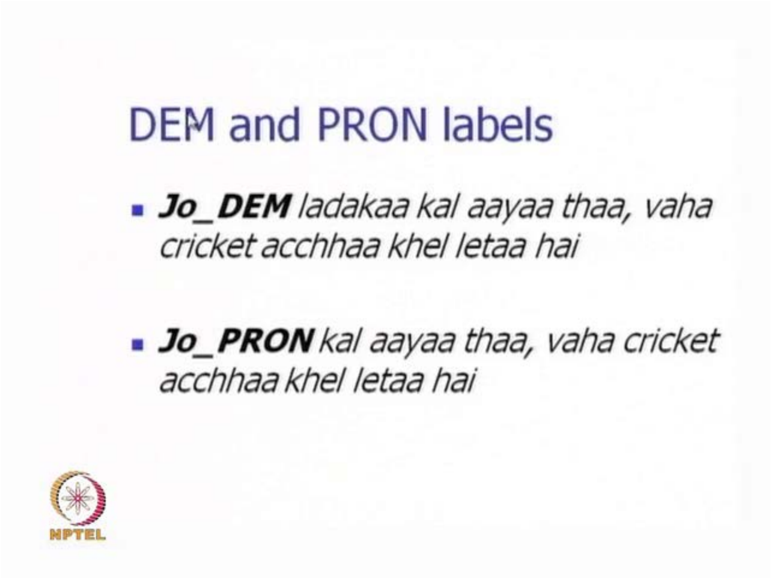


Disambiguation rule-1

- ***If***
  - ***Jo is followed by noun***
- ***Then***
  - ***DEM***
- ***Else***
  - ***...***

So, we formulated this disambiguation rule, which is pretty of S, if Jo is followed by noun, then it is a demonstrative.

(Refer Slide Time: 31:07)



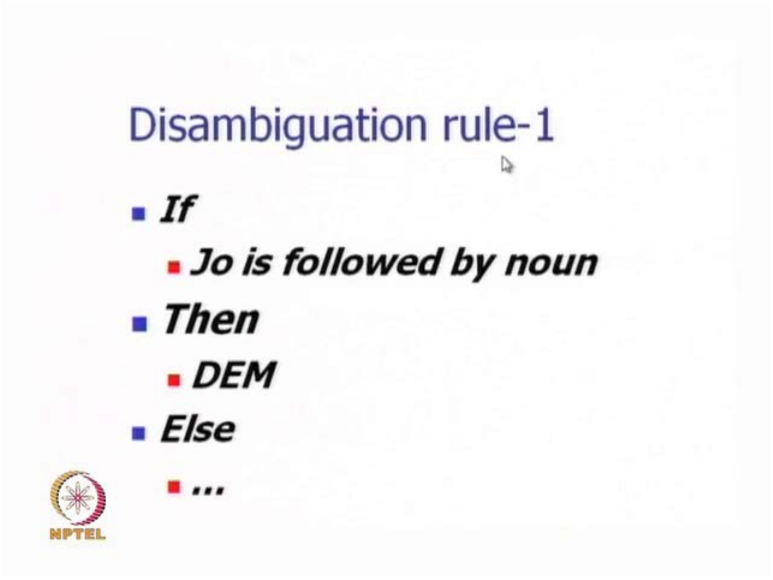
DEM and PRON labels

- ***Jo_DEM*** *ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

- ***Jo_PRON*** *kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

So, you can see the previous transparency that, Jo was followed by noun here. And therefore, it is a demonstrative.

(Refer Slide Time: 31:16)

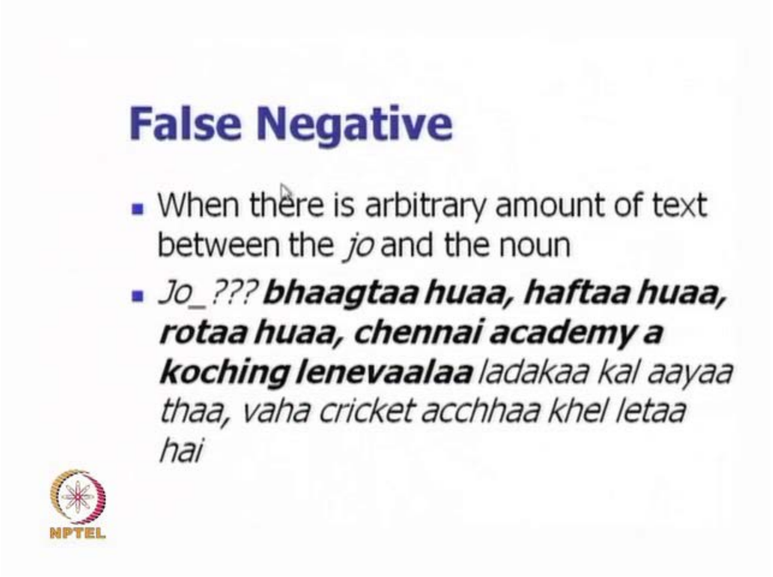


Disambiguation rule-1

- ***If***
  - ***Jo is followed by noun***
- ***Then***
  - ***DEM***
- ***Else***
  - ***...***

You are possibly already saying some problems here and we will discuss those problems, so Jo is followed by noun, it is a demonstrative else, we have to take more complicated steps to find out what Jo is.

(Refer Slide Time: 31:29)



False Negative

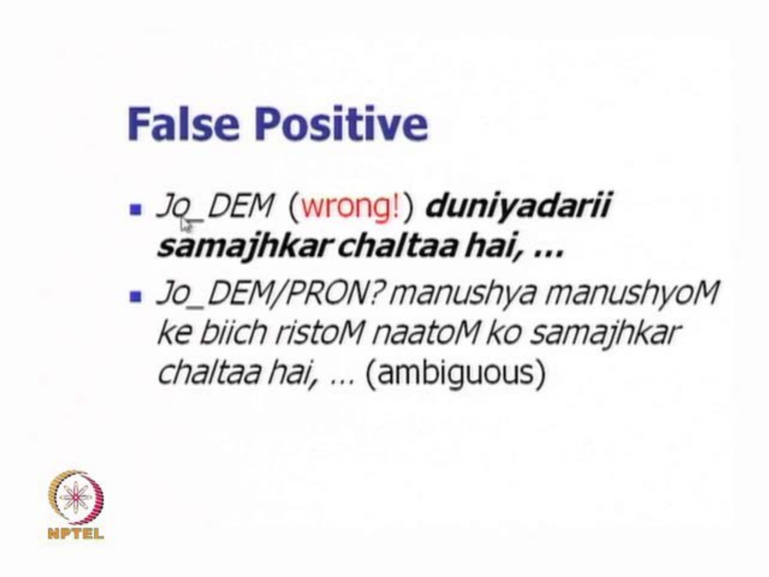
- When there is arbitrary amount of text between the *jo* and the noun
- *Jo_???* ***bhaagtaa huaa, haftaa huaa, rotaa huaa, chennai academy a koching lenevaalaa*** *ladakaa kal aayaa thaa, vaha cricket acchhaa khel letaa hai*

Now, we have the problem of what is called false negative and what is called false positive for any rule, which is suppose to produce a level or for that matter for any rule, it is possible to get into false negative and false positive situations. Let me illustrate from this example itself when, there is arbitrary amount of text between Jo and the noun, than the rule that, we have formulated, we will fail. So, take the sentence here for a movement forget about this or ignore, this piece of text in capital for a movement, if you ignore this then we have the sentence Jo [FL] cricket [FL].

So, this Jo and [FL] Jo is clearly demonstrative, the problem is that, Jo will continue to be demonstrative, even when it is not followed by a noun, because you have a an arbitrary amount of text between Jo and [FL], this text is seen here, [FL] Chennai academy may coaching [FL], this should be may Chennai academy may coaching [FL] cricket [FL] ok. So, sentence is a slightly artificial one, what an interesting sentence still Jo [FL] a Chennai academy may coaching [FL] cricket [FL] that means, the boy who came running and who was panting [FL] and who was crying [FL] and who was taking coaching in Chennai academy, he goes to a cricket coaching class this boy, who came yesterday he plays cricket well.

So, all this are modified, so are the qualifiers for this word [FL], so Jo [FL] Chennai academy a coaching [FL] cricket [FL]. So, again Jo is a demonstrative, but see Jo is not followed by noun [FL], this is a verb and therefore, this rule will fail and this is a case of false negative, we are not able to place demonstrative on this Jo, using that rule. So, this is a case of false negative, it is saying most probably the tag is not DEM and is failing and therefore, this is failure is not desirable ok.

(Refer Slide Time: 34:24)



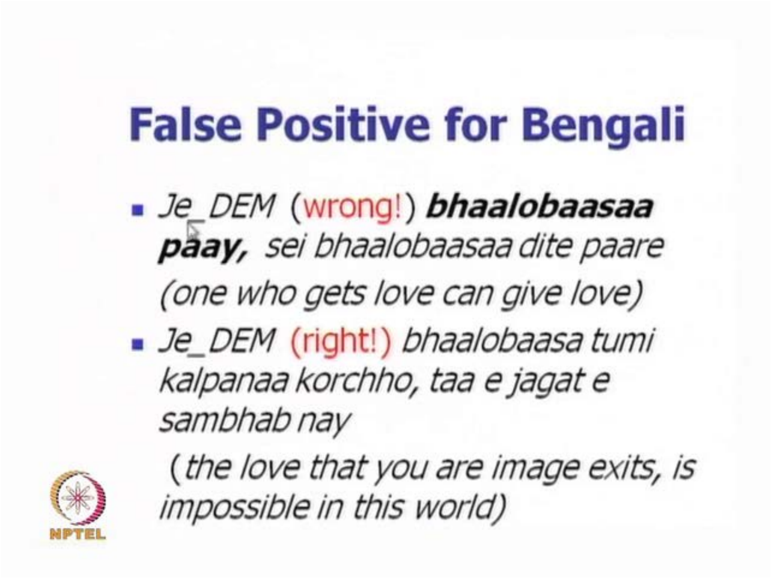
False Positive

- *Jo_DEM* (wrong!) ***duniyadarii samajhkar chaltaa hai, …***
- *Jo_DEM/PRON? manushya manushyoM ke biich ristoM naatoM ko samajhkar chaltaa hai, …* (ambiguous)

So, this is a case of false negative and there can be case of false positive for example, take this sentence here, Jo [FL] say this is the sentence, one who understands the ways of the world achieve success. So, in this case, if we place the demonstrative tag will go wrong, because the rule says that, if Jo is followed by a noun [FL] is noun here. So, it follows JO and therefore, it has placed simple mindedly a DEM the demonstrative and it is gone wrong. So, this is what we mean by false positive, this is wrongly given the demonstrative tag, this is an interesting sentence where, demonstrative of pronoun tag cannot be decided, because of ambiguity. So, the sentence is Jo [FL] and then there is a piece of text, so Jo [FL] the it has 2 meanings, one meaning is [FL] here is person, the person who understands the relationship between human beings [FL].

So, the person who understands the relationship between human beings can be called a compassionate person for example, this is a sentence. So, in this case [FL] is demonstrated by Jo and therefore, this can be the this can be given the level DEM, but see another reading for this sentence Jo [FL] here [FL] means man and men relationship between man and man. So, they go together and this Jo has an un specified noun, which it refers to, so this JO therefore, will be pronoun for that reading, now this sees a very difficult problem at the level of past again, you can out resolve this issue, because it requires grouping together [FL] or living out that grouping, this [FL] is separate from [FL]. So, this shows a case of false positive where, Jo can be wrongly given the tagged them and in this case, it is difficult to decide, it can go either way. So, therefore, a simple rule like this where, Jo followed by noun should be given the DEM tag is very brital, it can have both false negative and false positive.

(Refer Slide Time: 37:33)



False Positive for Bengali

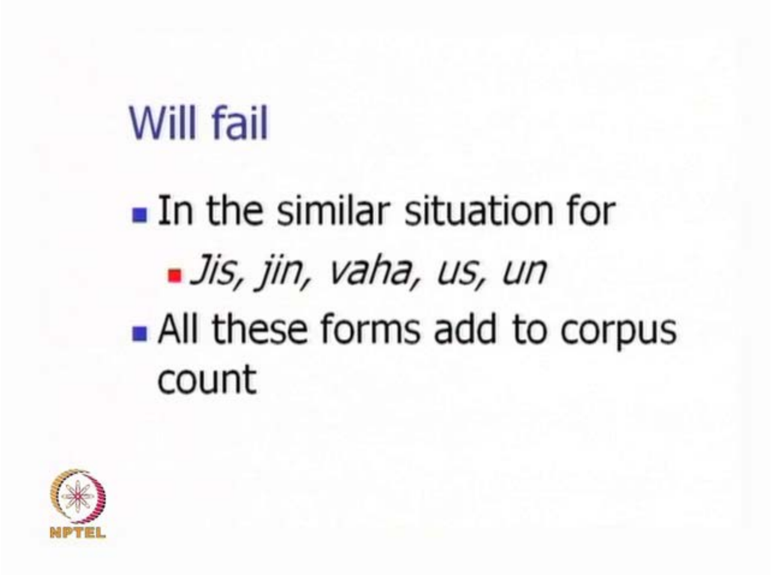
- *Je_DEM* (wrong!) ***bhaalobaasaa paay,*** *sei bhaalobaasaa dite paare* *(one who gets love can give love)*
- *Je_DEM* (right!) *bhaalobaasa tumi kalpanaa korchho, taa e jagat e sambhab nay*
  (the love that you are image exits, is impossible in this world)

We take the case of Bengali where, morphology markings are quiet weak on the words and we find here interesting cases, the sentence here is Je [FL], so that means, ones who gets love can give love. So, in this case again, the rule that, if Jo Je equivalent in bengali Je is followed by noun then it should be given the DEM tag, this goes wrong here [FL] is a noun it is after Je, but it will go wrong, because this Je is not referring to [FL] here.

This Je is referring to an unspecified noun not in the sentence, so this wrong and in this case however, the rule is working right. And here, we have the sentence as [FL] the love that, you imagine exits, this needs a bit of correction the love that, you imagine exits is impossible in this world Je [FL] here, this Je is the demonstrative for [FL] and therefore, if you give the tag them it is right. So, we have a very interesting situation where, the word J is followed by identical noun, but the further sentential context shows where, them would be wrong and where it will be right.

So, this shows it is a difficult problem, let us have a bit discussion on this now. So, what is happening is that, this was like Jo Je, they have both demonstrative role and pronoun role and this clue for whether, it is a demonstrative or pronoun can come from far apart in the sentence. So, the clue can be far away or the clue can be, because of some kind of syntactic structure, the sentence that we had in Bengali [FL] pi. So, in this case the fact, that Je is not demonstrative for [FL], comes from the word pi Je pi 1 who gets. So, this is little far away from Je and it can be quiet far away depending on arbitrary amount of text being inserted and therefore, this disambiguation would be difficult.

(Refer Slide Time: 40:31)

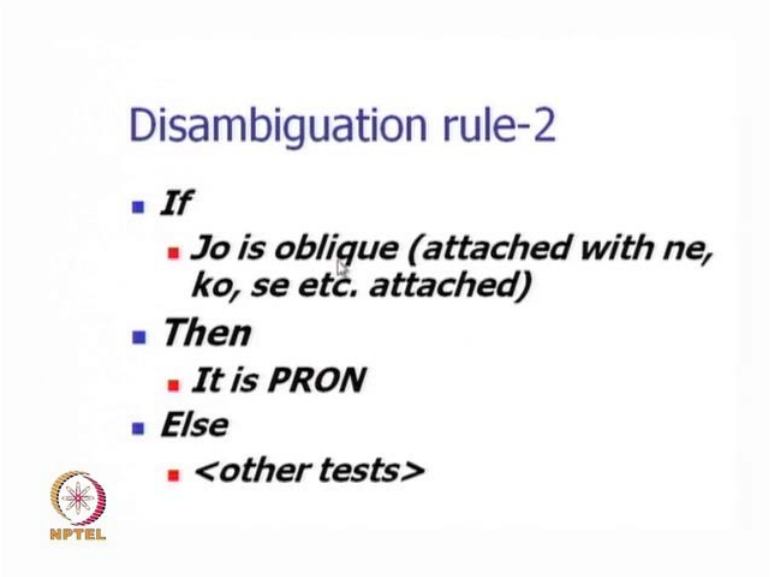


Will fail

- In the similar situation for
  - *Jis, jin, vaha, us, un*
- All these forms add to corpus count

So, we proceed further, we see that other forms of Jo like Jis Jin [FL] us un, they can fail in similar situation and all this forms are very very frequent, in the corpus. Therefore, all this put together can lead to a large amount of error of pos tagging all this errors can accumulate and you could have a situation where, the accuracy is pretty low, because simple rules cannot disambiguate the situation.

(Refer Slide Time: 41:07)

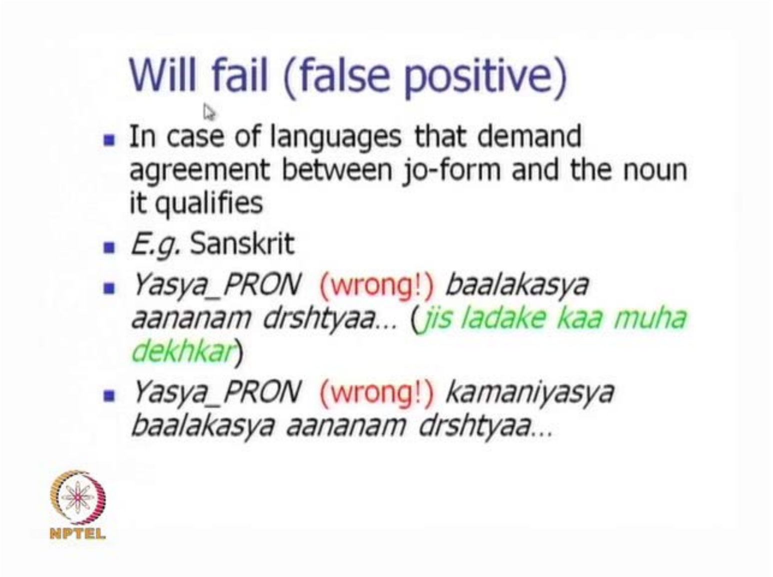


Disambiguation rule-2

- ***If***
  - ***Jo is oblique (attached with ne, ko, se etc. attached)***
- ***Then***
  - ***It is PRON***
- ***Else***
  - ***<other tests>***

We take the another disambiguation rule, rule number 2, which says that, if Jo is oblique it is attached with ne Ko se etcetera. If Jo is oblique then it is pronoun. So, this also looks like and accurate rule, if we do not examine this closely, we may get in a impression that, this is all there is in giving the pronoun tag to Jo. So, let us repeat this rule once again, if Jo is oblique that means, it is associated with some case mark and ne Ko se etcetera, then it is pronoun else, there are other tests.

(Refer Slide Time: 41:53)



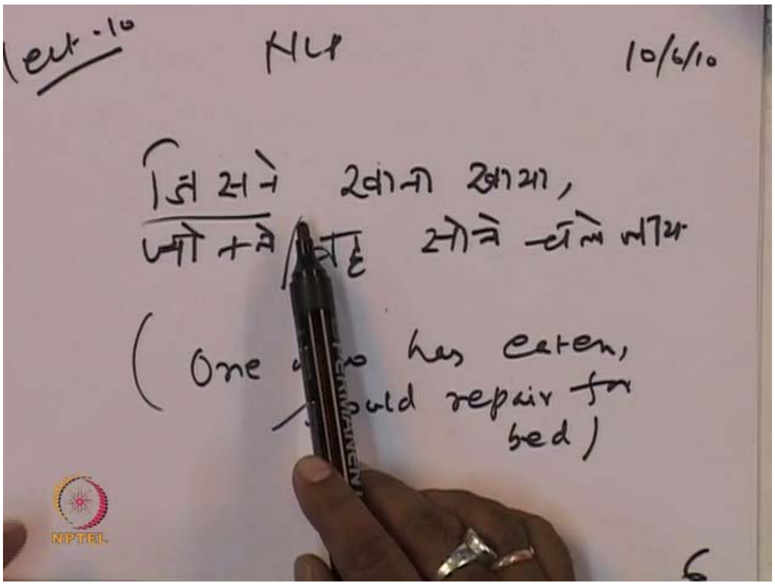
Will fail (false positive)

- In case of languages that demand agreement between jo-form and the noun it qualifies
- *E.g.* Sanskrit
- *Yasya_PRON* (wrong!) *baalakasya aananam drshtyaa*... (*jis ladake kaa muha dekhkar*)
- *Yasya_PRON* (wrong!) *kamaniyasya baalakasya aananam drshtyaa*...

So, this will fail, this will have false positive in case of languages that, demand agreement between Jo form and the noun, it qualifies for example, in [FL]. So, this is the meaning of this sentence Sanskrit insists that, the demonstrative and the noun it is demonstrative for must have forms that agree. So, here it is sasti vibakthi on baalakasya of the y, so [FL] has to be on the Jo for mace also [FL].

But, in this case, this rule that, if the Jo form is in case marked form, than it is necessarily pronoun that rule will fail, now the example is with Sanskrit, however it can very well hold for languages, which insist on this kind of agreement, we have another case here [FL]. So, here come on [FL], which means beautiful, which is qualifier for [FL], indicates that between [FL] and [FL] there can be arbitrary amount of text leading to the rule application going wrong. So, this indicates the complexity of formulating a rule. Now, may be a clarification is required here as to what, we mean by this oblique form and why is the case that, it will definitely get the pronoun tag, let us take an example.

(Refer Slide Time: 44:00)



Lect-10 NLP 10/6/10

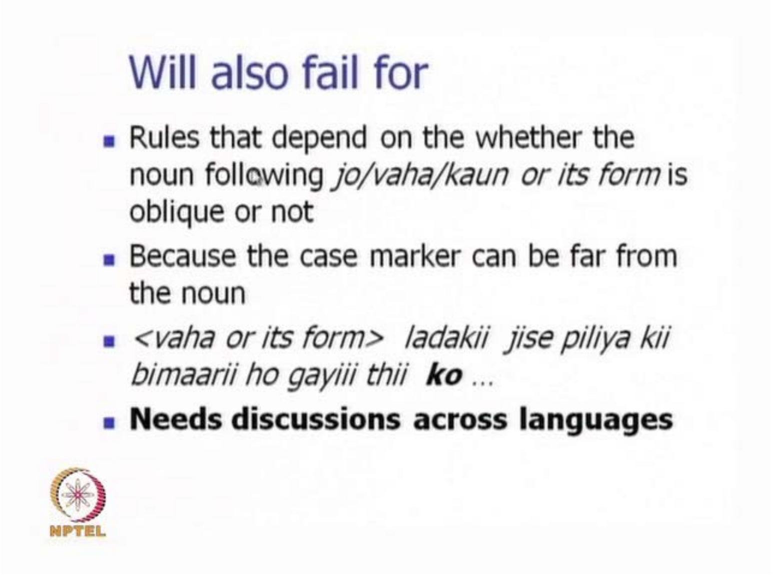
जिसने खाना खाया,
जो + ने वह सोने चला जाय

(One who has eaten,
should repair for
bed)

6

This is the example I will take, so we take the sentence [FL] this [FL], so one who has eaten should repair for bed, so one who is eaten should go to sleep [FL]. So, this is what, we mean by oblique form Jisneh is the oblique form of Jo, this is Jo [FL], now in Hindi the oblique form of Jo will always be a pronoun form. So, that is why that rule is quiet safe in Hindi, it is hundred percent accurate [FL] here, this will be a pronoun, but we are saying that, when there are languages, which insist on the [FL] coming on the Jo form well, it is demonstrative for a noun, which insist on agreement between the Jo form and the noun it is demonstrative for. In such cases, this rule will go wrong and we have seen an example in Sanskrit where, this kind of situation holds and other languages also may have this kind of agreement demand.

(Refer Slide Time: 45:41)

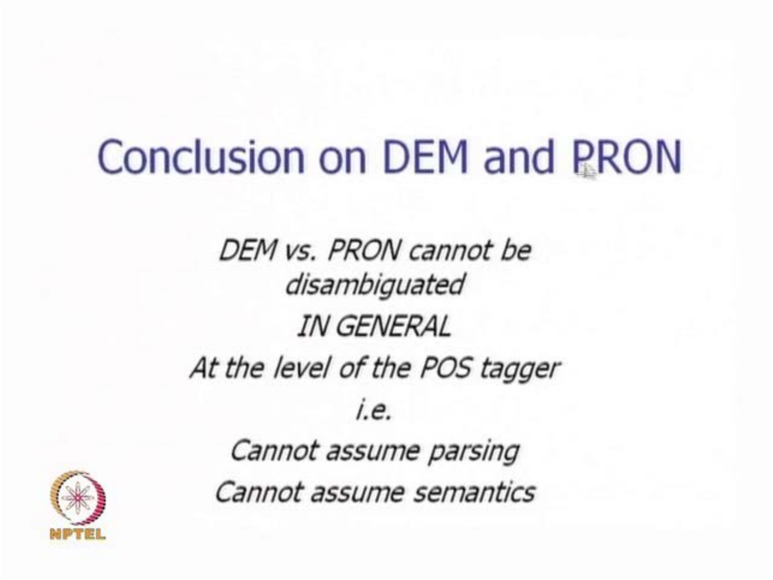


Will also fail for

- Rules that depend on the whether the noun following *jo/vaha/kaun or its form* is oblique or not
- Because the case marker can be far from the noun
- *<vaha or its form> ladakii jise piliya kii bimaarii ho gayiii thii* ***ko*** *...*
- **Needs discussions across languages**

Now elaborating further on this rules that depend on whether the noun following [FL] or it is form is oblique or not can also fail, because the case marker can be far from the noun. So, we can have constraints like [FL] or its form [FL], so the girl who was afflicted with jaundice the girl, who have jaundice is the meaning of this part of this sentence [FL] and this kind of construction is common in Hindi these days. So, the case marker is quiet far from [FL] and that is why, we need to be a careful here. So, we need to discuss phenomena across languages to see what kind of fails safe rules can be used for the demonstrative pronoun disambiguation and it is not a simple problem honesty many language evidences.

(Refer Slide Time: 46:56)



Conclusion on DEM and PRON

DEM vs. PRON cannot be disambiguated
IN GENERAL
At the level of the POS tagger
i.e.
Cannot assume parsing
Cannot assume semantics

So, the conclusion from this discussion on DEM and pronoun is that, DEM verses pronoun cannot be disambiguated, in general at the level of the POS tagger. That is we cannot assume parsing, we cannot assume semantics and if such clues are not present such information is not present then the demonstrative verses pronoun cannot be disambiguated in general, so that is the conclusion from this discussion on DEM and pronoun.

Now, one should at assume that, this is the only difficult case for part of speech tagging, there are many other such levels, which frequently get confused with each other one of them is main verb and auxiliary verb for India languages. It is also possible to confuse between noun and adjective, so if we have a an expression like golf clap or cricket bat here both cricket and bat are noun, but the first cricket is having an adjective function, it is an adject I will, which bat or what kind of bat cricket bat. So, here that is confusion between noun and adjective in the next class, we will discuss the mathematics of part of speech tagging and the algorithm.