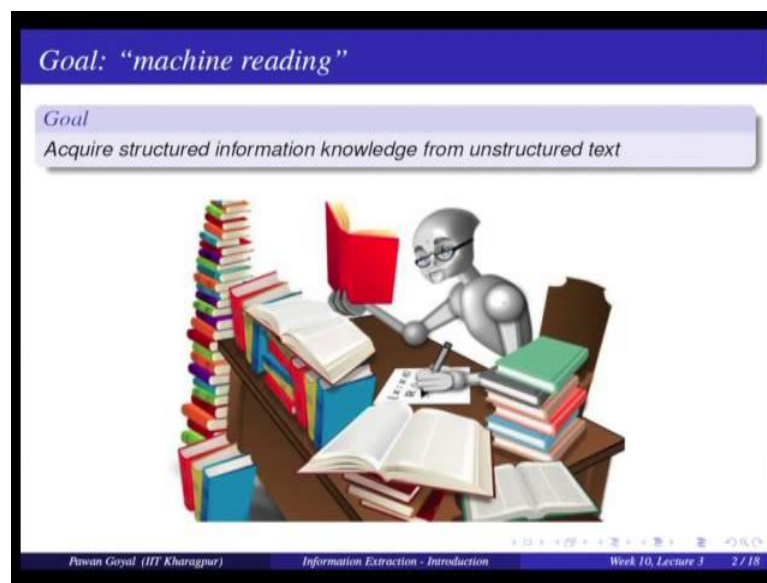


**Natural Language Processing**  
**Prof. Pawan Goyal**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology, Kharagpur**

**Lecture - 48**  
**Information Extraction – Introduction**

So welcome back for the third lecture of this week. So in this week we are doing some advanced topics on text mining. So this lecture we will start with information extraction. So we will see what are the basic, so all the basic applications where information extraction will be used, what kind of techniques you can use and we will focus our attention to a specific task that is relational extraction, how do I find out relation between any 2 entities by using the text corpus on the web.

(Refer Slide Time: 00:49)



So what do I mean by information extraction? So we can say that the goal of information extraction is like machine reading. So you have a lot of text available on the web. It is all on the unstructured form, there is no particular structure to that now from that text can I obtain some structured knowledge that can be used for various different applications and tasks. So this is like some sort of caricature to denote that that yes we have a lot of knowledge available on the web and machine is trying to go through the knowledge and get some a structured content that can be useful.

(Refer Slide Time: 01:29)

The slide is titled "Information Extraction" in a blue header. Below the header, there are two main sections: "Information Extraction (IE) Systems" and "Goals".

**Information Extraction (IE) Systems**

- Find and understand limited relevant parts of texts
- Gather information from many pieces of text
- Produce a structured representation of relevant information:
  - Relations (in the database sense)
  - A knowledge base

**Goals**

- Organize information so that it is useful to people
- Put information in a semantically precise form that allows further inferences to be made by computer algorithms

At the bottom of the slide, there is a footer with the text: "Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 3 / 18".

So what the information extraction systems do? They try to find and understand various different relevant parts of the text data. So what you will have. You will have a lot of text and you are trying to get certain important information from there. So we will see what are the various ways in which you can gather this information.

So from this all this data that is available, what you want, you want to get a structured representation of some sort of relevant information. And it can be like various relations in the sense of database. So you can find out what are the various entities involved in this text and what are the relations between those entities. And can also be some sort of knowledge base that you are constructing from the data.

So the goal of information extraction is that we organize information. So that it can be useful for 2 people for doing some of their tasks. And we want to put information in a very precise form that will allow need to make further inferences. So remember this was one of the things that we are talking about in the introduction also, that the natural language text is not very precise. So how do you make, how do you convert the information to something precise that can be used for doing various task, and various inferences. And that is what we doing in the case of information extraction.

(Refer Slide Time: 03:06)

**Information Extraction (IE)**

*Definition*  
Information extraction is the task of finding structured information from unstructured or semi-structured text.

*What sort of information?*  
IE Systems extract clear, factual information  
● Roughly: *Who did what to whom when?* etc.

*E.g., Gathering earnings, profits, headquarters etc. from company reports*  
● The headquarters of BHP Billiton Limited, and the global headquarters of the combined BHP Billiton Group, are located in Melbourne, Australia.  
● *headquarters("BHP Billiton Limited", "Melbourne, Australia")*

Pawan Goyal (IIT Kharagpur)      Information Extraction - Introduction      Week 10, Lecture 3      4 / 18

So this is a simple definition that you can get this is a sort of working definition you will say. So what is information extraction? So that is task of finding structured information from unstructured or semi structured text. So, you have the corpus or data that you have is either completely unstructured, so I can think of various treats various quora questions answers and lot of web pages' sort of unstructured. Or it can be somewhat semi structured where you have some more information like extraction information headlines etcetera, but this is not completely structured. So from this unstructured or semi structured data I want to find out a structure information, and that is what it is the definition of my information extraction.

So now, question comes in that what sort of information do you want to extract from here. So information can be something very clear and factual, like for example, this is this is something that that is normally done. So they should like who did what to whom when and who is in what relation to some other entity and so on.

So for example, suppose you have some newspaper text and they talk about various earnings profit headquarters etcetera. And this is one of the company report. The headquarters of BHP be Billiton limited and the global headquarters of the combined BHP Billiton group are located in Melbourne Australia. This is some sort of unstructured data that you can say in the form of the sentence.

Now, from this data you want to gather some structure information. So what kind of structured information do you get from this text? So you will see that, so what are the headquarters of BHP Billiton limited. So then you know the location here Melbourne Australia. And this can be some sort of structured information that you are trying to gather from the simple sentence. And that is what your information extraction system can do. So from here suppose you get this information headquarters of the BHP Billiton limited are in Melbourne Australia.

Since out of relation form there are 2 entities, and there is a relation between them and this you are extracting by using information extraction. So what is the use of doing this extraction? So once you do this extraction you will know all these tuples. So you will know these 2 entities are related are related by this relation so on and over there you can do lot of queries you can do lot of inferences and so on this can be very helpful for many question task also.

(Refer Slide Time: 05:52)

*Information Extraction (IE)*

*Example*

In 1998, Larry Page and Sergey Brin founded Google Inc.  
We can extract the following information,

- FounderOf(Larry Page, Google Inc.),
- FounderOf(SergeyBrin, Google Inc.),
- FoundedIn(Google Inc., 1998)

Such information can be used by search engines and database management systems to provide better services to end users.

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 5 / 18

Another example let us say we have the sentence, in 1998 Larry page and Sergey Brin founded Google.

Now, from this sentence what kind of information you can get. So who are the founders of Google, and when they find when they found Google? So all this information can be extracted from here and put in a very a structured form. So like I can have a information like founder of Larry page Google founder of Sergey Brin Google and founded in

Google in 1998. So all this information is there in this text and this can be extracted by a using information extraction systems.

So now once you have this information it can be used by various search engines and database management systems to provide better services to the end users. It is not very trivial to do it directly by using the text data, but once you have this in the database form you can do a lot of different tasks and look you can use a lot of different tools to make each of this information.

(Refer Slide Time: 07:00)

The slide is titled "Applications in Biomedical domain" in a blue header. Below the header, there is a light blue box with the title "Biomedical domain" and a bulleted list of challenges. The list includes: "A large amount of scientific publications", "Need to look for discoveries related to particular genes, proteins or other biomedical entities", "Biomedical entities often have synonyms and ambiguous names", and a "Critical task" which is "automatically identify mentions of biomedical entities in text and link them to their corresponding entries in existing knowledge bases." At the bottom of the slide, there is a footer with the text "Pawan Goyal (IIT Kharagpur)", "Information Extraction - Introduction", "Week 10, Lecture 3", and "6 / 18".

*Applications in Biomedical domain*

*Biomedical domain*

- A large amount of scientific publications
- Need to look for discoveries related to particular genes, proteins or other biomedical entities
- Biomedical entities often have synonyms and ambiguous names
- **Critical task:** automatically identify mentions of biomedical entities in text and link them to their corresponding entries in existing knowledge bases.

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 6 / 18

So now what are the various applications of information extraction? For example, biomedical domain, so in biomedical domain you have a lot of research papers that are published that give details about what about the various experiments that were done using and using various patients, and what was the findings of those experiments, and what kind of drugs work what kind of drugs does not work. They can be various clinic clinical trials they can be various patrons and all that lot of information is there, but this is already very unstructured form.

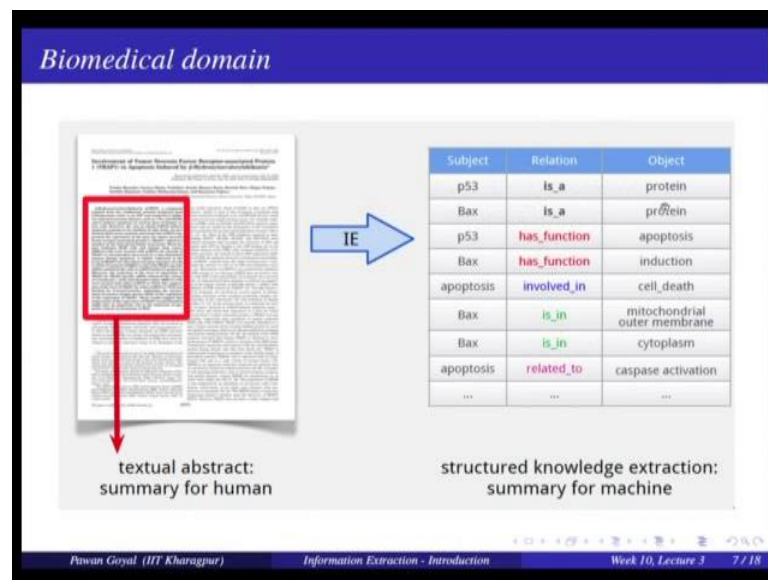
So suppose I need to look for discoveries that are related to various genes, proteins or other biomedical entities. And then the problem here could can be, that these entities can have various synonyms and there are lot of ambiguities involved. So what is the task? I need to automatically identify what are the mentions of biomedical entities in the text. I

find out these are that entities that has mentioned in the text. And then I want to link them to their corresponding entries in the lexical database.

Suppose I have a database that says these are all the different biomedical entities. Now in a research paper I need to find out this entity talks about this is corresponding to the particular entity in the database. This is very similar to the entity relenting problem that we discussed in this week itself.

Now, once we find out various entities in in the document, other task here could be that I want to find out how they are related to each other. So this is called relational extraction. That is one of the focus of the next 3 lectures.

(Refer Slide Time: 08:43)



So this is an example. So you have this research paper in biomedical dome domain and this is research paper you also get some abstract. Now from this abstract can you extract information in a structured format, like p53 is a protein, bax is a protein p53 has function of apoptosis so on. Now all this information is available in the text data, but not in this very nice structured format. So from there can you extract these are the entities and this is the relation between them.

So you find what are the entities and with different between various pairs of entities what is the relation. And this is called the structured knowledge extraction, and this is the analogy is shown here. So the research paper extract can be thought of as if something

for humans and this structured knowledge means can be thought of as something for machines. So machines can make use of this information for various tasks now.


(Refer Slide Time: 09:46)

*Relation Extraction*

CHICAGO (AP) — Citing high fuel prices, United Airlines said Friday it has increased fares by \$6 per round trip on flights to some cities also served by lower-cost carriers. **American Airlines**, a unit of **AMR**, immediately matched the move, spokesman **Tim Wagner** said. **United**, a unit of **UAL**, said the increase took effect Thursday night and applies to most routes where it competes against discount carriers, such as Chicago to Dallas and Atlanta and Denver to San Francisco, Los Angeles and New York.

Subject	Relation	Object
American Airlines	subsidiary	AMR
Tim Wagner	employee	American Airlines
United Airlines	subsidiary	UAL

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lec 1



Another example: so this is like a report that that you find on the web, and from this report can you extract various relations. So here you have the sentence American airlines the unit of AMR immediately matched the move a spokesman Tim Wagner said. So from here you can find out the Tim Wagner region is a spokesman for American airlines, and suppose your relation is employee. So we can say Tim Wagner is employee of American airlines, also American air airlines is the unit of AMR. So you can have this relation American airlines is a subsidiary of AMR, and similarly here united a unit of UAL you can find out this relation again.

So from this huge amount of text data can you find out this is structured information. So that is the task of information extraction. Find out the entities and what are the relations between them.



(Refer Slide Time: 10:45)

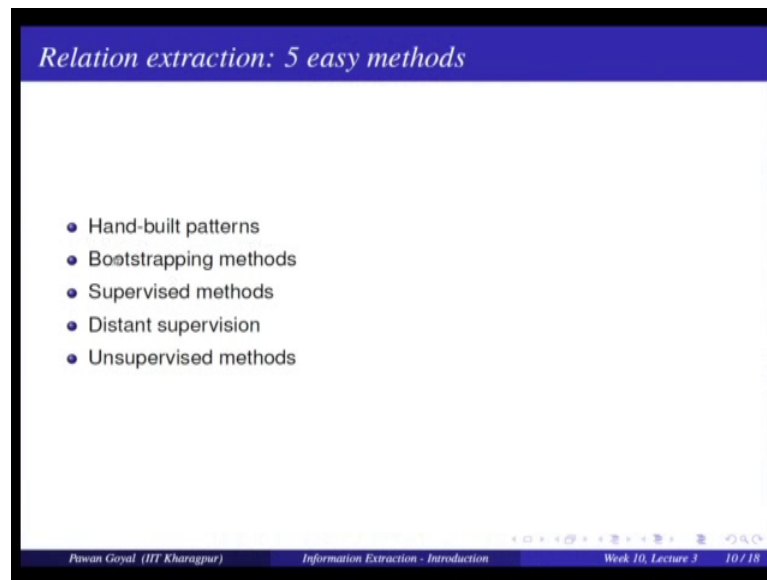
Relation types			
For generic news text ...			
Relations	Examples		Types
Affiliations	Personal	<i>married to, mother of</i>	PER → PER
	Organizational	<i>spokesman for, president of</i>	PER → ORG
	Artifactual	<i>owns, invented, produces</i>	(PER   ORG) → ART
Geospatial	Proximity	<i>near, on outskirts</i>	LOC → LOC
	Directional	<i>southeast of</i>	LOC → LOC
Part-Of	Organizational	<i>a unit of, parent of</i>	ORG → ORG
	Political	<i>annexed, acquired</i>	GPE → GPE

Another example: so when the relation can be were also very generic. So like you can be personal relations like married to mother of organization relations like a spokesman for president of artifactual owns, something invented something, produces something they can geospatial relations, that this city is near to this city, in the on the outskirts of the city. And these kind of relations might be very helpful in replying to various queries they talk about that need geography information. See you know what cities are nearby other city. So you can try to answer these kind of questions. And directional relation this is southeast of So on. And they can be part of relations. So you need of something parent of annexed acquired for this political relation.

So you can think lots of different relations that can be established between entities. Now using these relations, you can do lot of different some sort knowledge engineering, you can do a lot of inferencing you can try to answer questions and try to predict certain relations between the entities there are lot of different tasks that you can use once do you can do once you have this structured information.



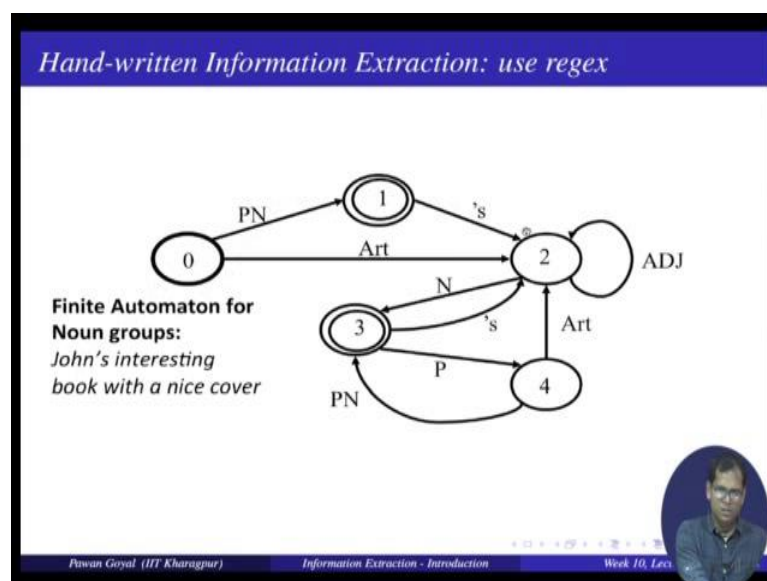
(Refer Slide Time: 12:06)



So now, so the topic here is how do we gather this research information. So there are like I would say like in many different NLP applications. So here also there are 5 different methods for doing this task. So one simple method is you choose your hand built patterns. Then you get it bootstrapping methods you know supervised methods distance supervision is a very nice idea that you will see for this particular task and then you can also use some unsupervised methods.

So we will focus on the first 4 methods and we will see clearly how you can use one of these methods for the task of information extraction.

(Refer Slide Time: 12:46)



So let us see what we do in the hand built patterns. So idea is you can use various regular expressions for finding entities and the relations between them. So suppose you want to find the entities. So this is for noun groups. It is simple regular expression. So regular expression you can also denote by using a finite automaton. So here you are seeing a finite automaton that is denoting a regular expression, and this so what it is denoting any noun group.

So let us try to follow this. So you have this phrase John's interest interesting book with a nice cover. This is a noun group. So how does this automatic capture this? You say a pronoun a personal noun John, John's interesting becomes an adjective. Book is a noun with is a preposition, article nice adjective cover noun and this is a it is a final state. So this becomes a noun group.

So we will see even John is a noun group. And John's interesting book is also a noun group. So it is trying to capture nouns group, noun group in various sort of granularity. You can even have single word you can have multiple words. So you it is telling you what is a noun group. Now you can further extend it to find out I know what are the noun groups now what is the relation between that.

(Refer Slide Time: 14:22)

*Rule-based Extraction Examples*

*Determining which person holds what position in what organization*

*[person], [position] of [org]*  
Vuk Draskovic, leader of the Serbian Renewal Movement

*[org] (named, appointed, etc.) [person] Prep [office]*  
NATO appointed Wesley Clark as Commander in Chief

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 12 / 18

So suppose I want to find out which person holds what position in what organization. So what kind of patterns I can think of. A person x holding position y in organization z, so suppose I have to use some hand built patterns how will I go about it, so I will first think about what are the various kind of sentences where all these 3 entities can occur together. So there will be a person who is working in an organization. So and then once I have found some sentence is I will try to abstract what is the normal pattern that I am seeing here.

So for example one pattern can be person comma position of organization. Because you find sentences like vuk draskovic is a person comma position leader of the Serbian renewal movement. So now, what you are abstracting here, there is a person position and organization. And now you can think of many such sentences, where all these 3 entities will be there in this relation. So once you identified this pattern you will give this pattern to the machine and from there corpus it can extract all these entities for you and you will know immediately that these entities are connected by a particular relation.

What can be other patterns? So like organization named appointed etcetera person preposition office again they are all these entities. So NATO appointed Wesley Clark as commander in chief. So we are finding again all the 3 entities in a particular relation. So similarly suppose your task is to find out where is an organization located. So we will think about what are the patterns something like x located in y and or y is xs

headquarters. So we will think of these patterns and using these patterns you will try to extract these pairs of x y.

(Refer Slide Time: 16:32)

The slide is titled "Rule-based Extraction Examples" in a blue header. It contains two examples of rule-based extraction. The first example is titled "Determining where an organization is located" in a pink box. Below it, a blue box shows the rule "[org] in [loc]" and the example "NATO headquarters in Brussels". The second example is titled "[org] [loc] (division, branch, headquarters, etc.)" in a blue box. Below it, a blue box shows the rule "[org] [loc] (division, branch, headquarters, etc.)" and the example "KFOR Kosovo headquarters". At the bottom of the slide, there is a footer with the text "Pawan Goyal (IIT Kharagpur)", "Information Extraction - Introduction", "Week 10, Lecture 3", and "13 / 18".

Like organization location NATO headquarters in Brussels. So we will extract NATO headquarters and Brussels are the 2 entities. Organization location and you can say division branch headquarters like KFOR Kosovo headquarters. So you know this is the observation and it is a location.

So like that you can think of various patterns and extract these relations. And this is one of the very early examples on how these kind of patterns for used for extracting hyponym relation. So hyponym which is you remember, it is a relation between sub concept and a super concept. When these words first given by Hearst.

(Refer Slide Time: 17:11)

*Patterns for learning hyponyms*

*Intuition from Hearst (1992)*

Agar is a substance prepared from a mixture of red algae, such as Gelidium, for laboratory or industrial use.

- What is Gelidium?
- How do you know?

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 14 / 18

So what is the basic intuition? Suppose you are seeing this sentence agar is a substance prepared from a mixture of red algae such as gelidium, for laboratory or industrial use. This is sentence. Now suppose I ask what is gelidium, and you can say gelidium is some sort of algae or red algae from the sentence, yes; now how do you know that gelidium is a red alga? See you are seeing some sort of pattern here. Red algae such as Gelidium, so this pattern is telling you that gelidium is the kind of red algae.

Now, you can try to abstract these pattern, in you say that whenever you are finding such patterns x such as y, there is a hyponym-hyponym relation between x and y. And this is the idea find out many such patterns and from these patterns you try to extract these entities. So what has did, he found out various search patterns where you can have 2 entities connected by hyponym relation.

(Refer Slide Time: 18:22)

### Hearst's lexico-syntactic patterns

#### Automatic Acquisition of Hyponyms

- Y such as X(,X)\* (, and/or) X)
- such Y as X
- X or other Y
- X and other Y
- Y including X
- Y, especially X

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 15 / 18

So y such as x is for hyponyms. So what are the other kind of patterns you can use such y as x like such. Vehicle as car such vehicle as bicycle x or other y yes car or other vehicle car and other vehicle, vehicles including car and so on, vehicle especially car. So this I am given example with car and vehicles, but you can think of it as with any hyponym-hyponym pair. So he found out Freddy such titles and from these patterns he tried to extract the hyponym, hyponym relation from the text data.

(Refer Slide Time: 19:05)

### Examples of Hearst patterns

Hearst pattern	Example occurrences
X and other Y	...temples, treasures, and other important civic buildings.
X or other Y	bruises, wounds, broken bones or other injuries...
Y such as X	The bow lute, such as the Bambara ndang...
such Y as X	...such authors as Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, including Canada and England...
Y, especially X	European countries, especially France, England, and Spain...

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 16 / 18

So here are some examples for these Hearst patterns and what kind of example occurrences you can see in the data. So the pattern x and other y, you can see temples tragedies and other important civic buildings. So from this sentence you can immediately see that 10 percent treasuries are sub concepts of civic buildings. So we can have this pair of hyponym-hyponym. Civic build civic buildings are the hyponym and temples is the hyponym. Similarly, treasure is the hyponym x or other y. So bruises would not broken bones or other injuries. So we can have all these as a hyponym of injuries y such as x. So the bow lute such as the Bambara ndang.

So here you can see that bow lute is the super concept this is the sub concept. Such y as x such authors as Herrick goldsmith and Shakespeare. So immediately you will see there is a relation here, so on y including x y especially x.

So Hearst Hearst manually found that all these patterns, and from these patterns he was trying to extract a hyponym-hyponym pair from the data.

(Refer Slide Time: 20:17)

*Patterns for learning meronyms*

*Berland and Charniak's patterns*

- Selected initial patterns by finding all sentences in a corpus containing *basement* and *building*

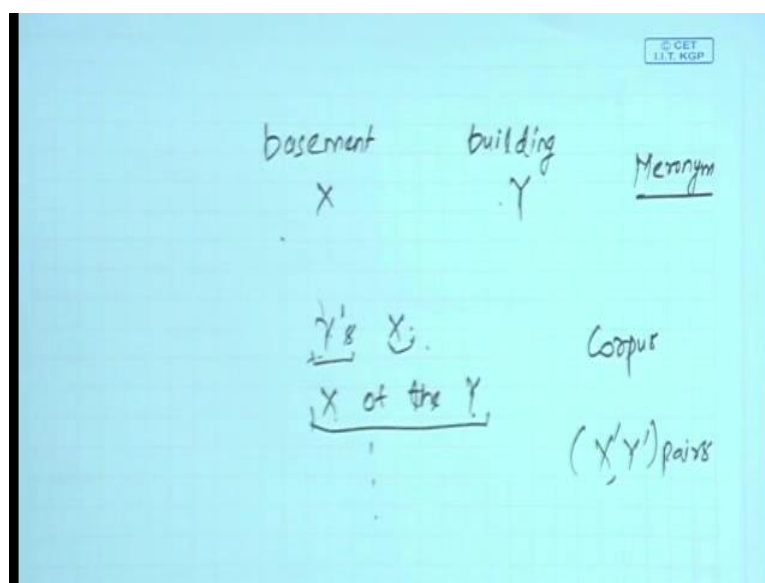
whole NN[-PL]'s POS part NN[-PL]	... building's basement ...
part NN[-PL] of PREP (the a) DET mods [JJ NN]* whole NN	... basement of a building ...
part NN in PREP (the a) DET mods [JJ NN]* whole NN	... basement in a building ...
parts NN-PL of PREP wholes NN-PL	... basements of buildings ...
parts NN-PL in PREP wholes NN-PL	... basements in buildings ...

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 17 / 18

Similarly, Berland and Charniaks, they found out some patterns for meronym relation, that is part of relation basement is the part of building. So they were trying to find out patterns for meronyms. So again you can think of what are the patterns that come to your mind. So like building's basement. So you will think of some example and see what kind of sentences they occurring building's basement of the building and so on and you will try to make patterns out of these.



(Refer Slide Time: 20:51)



So let us take these 2 simple examples. So like I am seeing that I have an example, basement and building. And this is my, suppose my X and this is my Y. And I want to find out many such X Y pairs that have the same meronymy relation. So how will I start? I say in the sentences how will basement in building occur together something like basements building sorry building's basement. So it will be Y's X building's basement or basement of the building X of the Y and so on and these are now my patterns. And then you will try to see in my corpus where do all these patterns occur. So example is cars wheel, wheel of the car. So we will see these X, Y are related by this meronymy relation and that is how you will try to. So we are in these patterns we will try together many such X, X prime Y prime pairs.

So what Berland and Charniak did? They selected some initial patterns for finding all sentences in the corpus that contain basement and building that is a normal is a nice method of finding these patterns. So then they found like building's basement, basement of a building basement in a building basements of buildings basements and buildings. So on now here they were writing down the patterns. So here something like NN. So they were writing in terms of what is the parts of speech that is coming and so on, so of preposition. So parts the plural noun of preposition wholes NN it is a plural noun.

So this is part in whole relation part coming as NN, in between there is the word in as a preposition or a as a determiner and in some modifiers. So there are now here

abstracting. So what they are seeing basement in a building, but it might be basement in a huge building right. So how do I absolutely better than I say there is in optionally they can be in adjective here. So that is why they are saying JJ or NN is star. Basement in a civic building and so on all these can be captured by slightly generalizing these patterns. So that is what you are seeing here JJ or NN. So you can have a civic building huge building and all this will be captured here.

So like that you try to find out these patterns and using these patterns. So once you have these patterns you try to extract some other entity pairs that are involved in this relation.

(Refer Slide Time: 23:41)

*Problems with hand-built patterns*

- Requires hand-building patterns for each relation!
  - ▶ hard to write; hard to maintain
  - ▶ there are zillions of them
  - ▶ domain-dependent
- Don't want to do this for all possible relations!
- Plus, we'd like better accuracy
  - ▶ Hearst: 66% accuracy on hyponym extraction
  - ▶ Berland and Charniak: 55% accuracy on meronyms

Pawan Goyal (IIT Kharagpur) Information Extraction - Introduction Week 10, Lecture 3 18 / 18

So this is a nice method if you want to sit down and look at each and every relation and think about the patterns. And that is also the problem with this approach that some persons who are good with the data, who are also language they know how the system work, they can they can try to get you some hand built patterns. So now, question the problem is that they are hard to write and hard to maintain and there are like you can think of zillions of patterns. So we can think of so many different ways in which people can talk about hyponym-hyponym pair in the data.

So how do I capture all of these patterns manually and yeah there might be domain dependent. So every domain you might have different ways of writing things, and yes you can do that for some relations, but suppose they are thousands of relations. How do you do for all these thousand relations? So we and these patterns that Hearst found or

Berland and Charniak found they were giving kind of results, but they were not like giving very accurate results. So for example, Hearst patterns give the roughly 66 percent accuracy on hyponym extraction, and Berland and Charniak gave 55 percent accuracy on meronyms.

So we would like probably prefer to have better accuracy than these numbers. So that is using hand built patterns you can only go little. So only small distance and there are also a lot of manual effort is required. So how can we avoid this manual effort? And that is what we will see in the other approaches, that we will be discussed it starting from, how do we do simple bootstrapping here, and that is all you will be discussing in the next lecture.

Thank you.