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

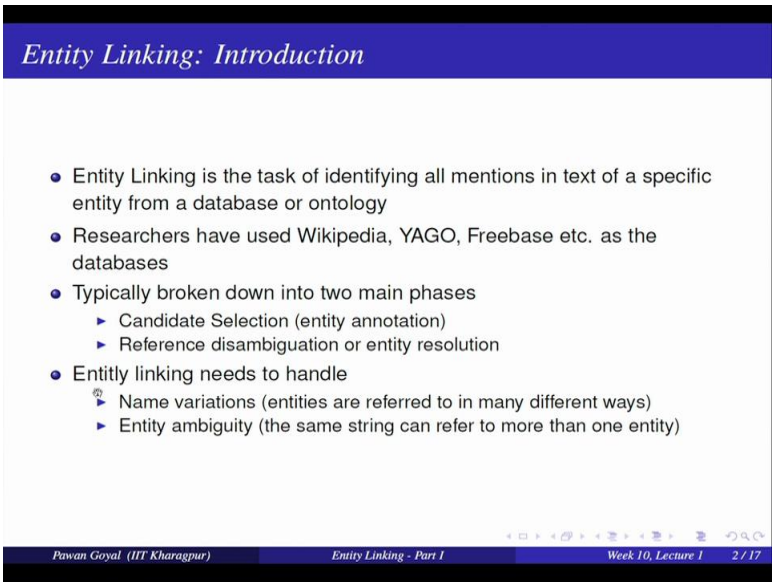
Lecture - 46
Entity Linking – I

Hello everyone. Welcome back to the week 10 of this course. So, we have been doing a lot of basics in this course and last week we finished our discussions on semantics. And during semantics we also talked about various applications where you can use topic models and other things like distribution semantics and (Refer Time: 00:35) semantics.

So, starting from this week we will focus mainly on applications. So, you will use both the basics that we have covered and some new methods for solving very specific tasks. And for this course we have chosen the tasks that are very very popular in the field of NLP and text panel. So, this week we will start with; so we will start the first two lectures which is entity linking and then we will go towards information extraction.

So, today we are starting with entity linking.

(Refer Slide Time: 01:06)



So, what is entity linking as such? What is the problem there? So we were discussing sometimes in the starting and in some lectures that when we encounter text data lot of entities are mentioned there. And it would be helpful for a task if you what are the

different entities that are used in this particular text data. Now there are various knowledge basis and resources where you have a good list of all these entities, and some descriptions of these available.

So, suppose you can find out from a text what are the important entities here and if you can link them to a knowledge base, then you can do a lot of further inferences over these. For example, this entity is a person and there in the in the database you will have ma lot of information about this and you can make use of all these information. So, this as if you are having some background knowledge about various entities given a text data when you encounter the entity you try to link it to the knowledge base and then you can extract the background knowledge about it. And use that for different task in this particular data.

So, this is a problem of entity linking. So, we can define it in this manner that entity linking is the task of identifying all mentions in text of a specific entity from a database or in ontology. So, what we are assuming here we have a database or we can also call it knowledge base ontology, where I know what are the different entities that I need. And with that inform with that entity I will have some different sort of information. For example in Wikipedia, think of all the Wikipedia pages you have, you can call, you can think of all these Wikipedia pages at the entity pages and you have lot of information about the entities in each page.

So, this is my knowledge base. Now when you encounter a text there your problem is find out if a particular n gram or a sequence of phrases sequence of words together correspond to an entity in the Wikipedia. And if so then link it to that Wikipedia page and this is the overall idea of entity linking problem.

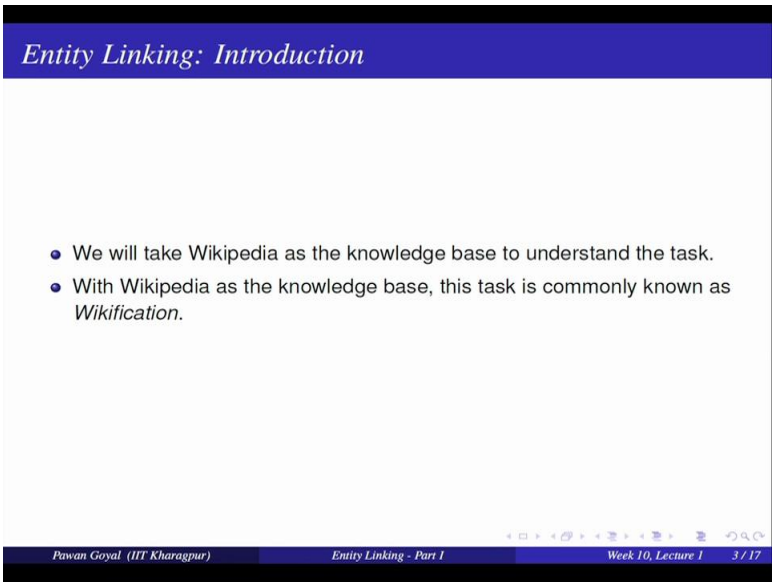
So, lot of different databases can be used. For example researchers have used Wikipedia lot then YAGO, freebase etcetera. And the task of entity linking when you are doing you can break it into two different phases. So, one phase would be you find out from the text what are the appropriate candidates or entities that should be linked. So, this is called the entity mention detection part. Identify what are the mentions of entities that should be linked to the database. And once you found out what are the important entities the next one would be to appropriately link it to the database that is the second part; reference disambiguation or entity resolution.

Now why would that be a problem? See same as we discussed in the case of words is disambiguation. With the same entity name there might be different reference. So, for example New York, it can be New York City and there might be movie with the name New York. There might be tv serial with the name New York. So, when in a text you have a mention of New York you want to know whether you want to link it to the New York City or the film or tv series or something else. So, there is a problem of disambiguation here. And this is also to be handled when we are solving the problem of entity linking. Find out what are the appropriate entities, and then appropriately link them to their entries in the database.

So, what are the things your challenge is that one needs to handle in this problem? So, one is mean variations. So, the same entity can be written in many different ways. For example, simple things like Hillary Clinton. So, in a text you can all you might write it with only the name Clinton with the last name Clinton or maybe there is a middle name also involved here. So, you can write it in different different manners. So, your problem is you have to handle all these name variations and all these should map to the appropriate entity in the knowledge base.

And then the second challenge is entity ambiguity; that is the same string can refer to more than one entity. So this is also we discussed, New York can refer to multiple entities. So, both these challenges are to be handled in the problem of entity linking

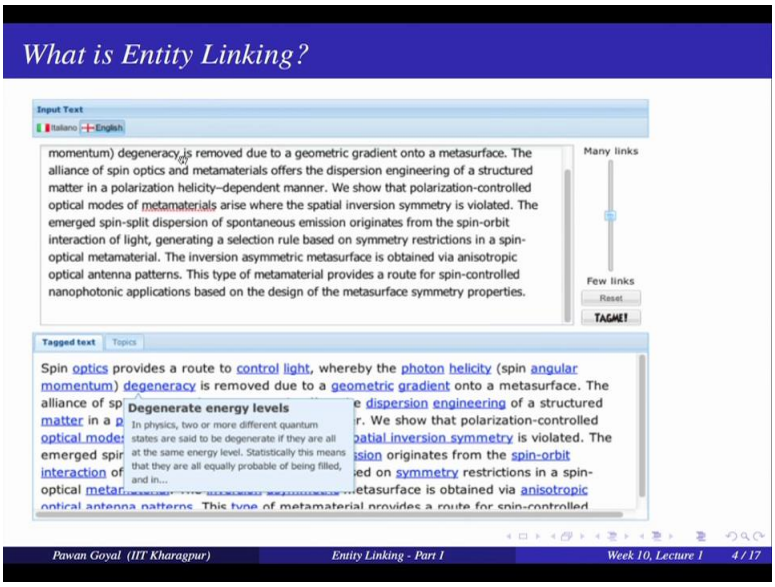
(Refer Slide Time: 05:37)



So now, for this course what we will do we will take one particular database that is Wikipedia as our base database. So, we will always link a text to Wikipedia. So, this will be our database by default for this lecture in the next lecture. But in general you can use any other database and you can accordingly modify your approaches for that. So one particular terminology; so if you are using Wikipedia as your knowledge base then this task of entity linking is called Wikification or Wikifying. So, you are taking a text and you are trying to Wikify that. That means, find out the entities that are important and then linked them to their appropriate Wikipedia pages; that is the problem of Wikification.

So, now let us going detail about this problem, what is the different processing involved and what are the different techniques you can used.

(Refer Slide Time: 06:31)



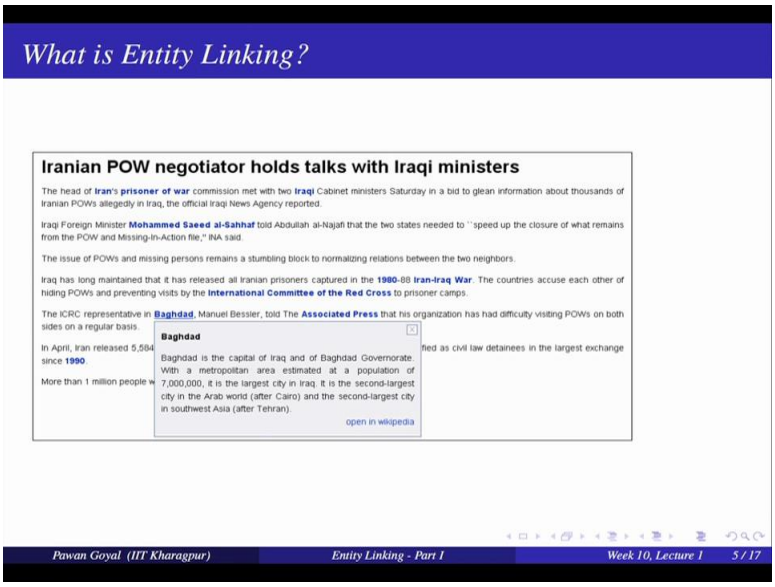
So, here is one example of entity linking or Wikification as such. So, on the top what you are seen, you are seeing a research article from physics domain. And you are having; so this like an abstract here, so you are having text such as degeneracy is removed due to a geometric gradient onto a meta surface and so on. So, there is a text involved here. Now what is the task you are trying to Wikify that? So that means, you are trying to identify what are the important entities and what are their pages in Wikipedia.

So, if you see on the bottom you are you are finding various links. So, spin optics. So, where optics is linked to some different page; an example is from degeneracy. So, this is

linked to degenerate energy levels that are a page in Wikipedia. and you can also find out some specific content about that page, if you just take your mouse over there. And that is being done to many different words here; optics, control, light, photon, helicity, angular momentum and so on.

That means, these words are the appropriate mentions and they are then linked to their Wikipedia pages. So, this is the process of Wikification.

(Refer Slide Time: 07:52)



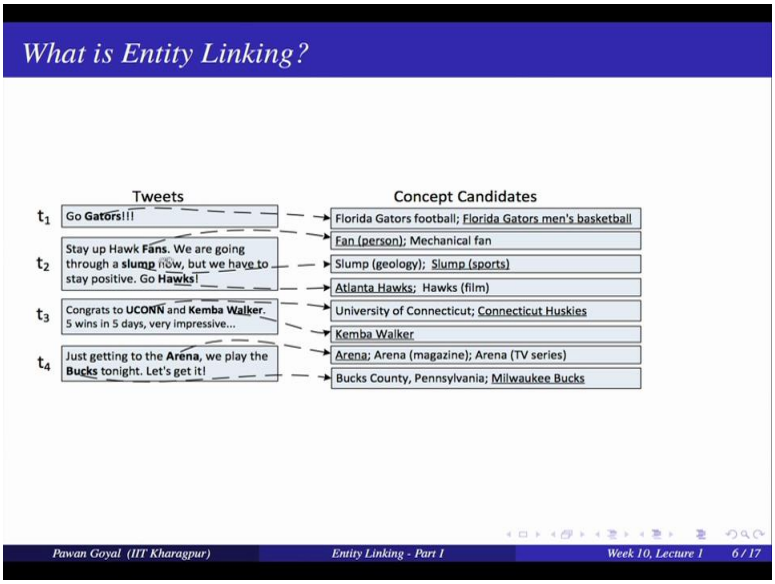
Similarly, here you are seeing a news article. In the article with the news you are seeing various words are hyperlinks. So, there you can click this word and go to their appropriate Wikipedia page. Also for example, here in Baghdad; so it will open the Wikipedia description of the word Baghdad and you might just want to read the summary or if you want to know more you can click on this opening Wikipedia and go to the Wikipedia page.

So, so we can also see why this might be very important application. So, you can talk about reading news articles or scientific articles also any other text it can be tweet text. So, when you reading the article there might be many different terms that are very very specific domain and if you are not in expert in the domain you will not know what these terms stand for.

So, then what you would do you will take this terms and try to search in the dictionary or some on Google or some or may be on Wikipedia, and that will require a lot of time from your side to fully understand that article. So, what is Wikification doing? It is helping you in that given a text it will automatically identify what are the important entities and it will link the Wikipedia pages. So, it will avoid, it will do all the tasks for you and you can even see the description in the same page or you can go to the Wikipedia page and read more about it.

So, that helps a lot in enhancing your reading for a certain news article or scientific article. Additionally it can also help if you are planning two certain tasks on that text. Your own to get some semantics from the text then also having this knowledge that this entity corresponds to this Wikipedia entity might help you in getting some knowledge from the Wikipedia page, and take it as a feature for your task. So, this is also one other application for this Wikification.

(Refer Slide Time: 09:55)



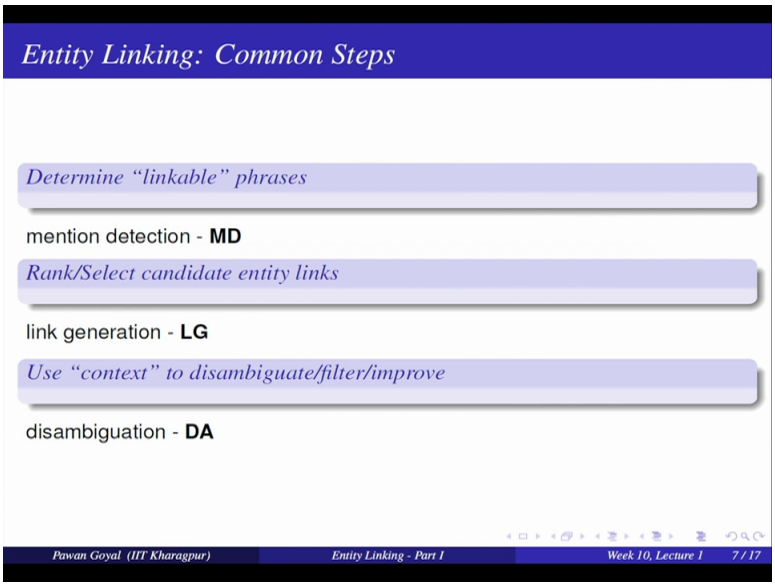
So, we have seen for scientific articles and news articles you can also do it for very short text like tweets. So, tweets have very little context. So, here you are seeing with four different tweets “Go Gators” and here the problem is what does this ‘Gator’ refer to. So, in Wikipedia there are two different mentions that are possible Florida Gators football of Florida Gators men’s basketball. In this particular context it refers to Florida Gators men’s basketball. And you want to link it to that particular entity.

Similarly here; “stay up Hawk Fans. We are going through a slump now, but we have to stay positive. Go Hawks”. So, here you have entities like fans slump and Hawks and they need to be linked to appropriate entities. Again you are seeing their multiple possibilities and you need to find out what is the appropriate mention Wikipedia; same with the other examples that you are seeing here.

So, you can do it for also for short text like tweets. With tweets there is a very little information and you want to find out what is the appropriate entity that this tweet links to. And you can also think of many other applications on your own where this entity linking is important.

So, now how is actually done? So for that let us try to understand what are the different phrases again, what do we need to do systematically.

(Refer Slide Time: 11:13)



So, what are the common steps? First step is it will determine what are the linkable phrases, and this step is also called Mention Detection. That is from the text find out what phrases what n grams are to be linked. For example, we were seen words like Baghdad and here Gators what the mentions that word to be linked to the knowledge base. And this approach for detecting these words is called mention detection.

Now once we have found out what are the appropriate mentions for to be linked then what will be the next step. Next step would be you have to identify what are the possible

candidates to which it can be linked. Like in the previous slide we were seen Gators can link to two different entries. So, identify what are the possible entries, and this is called the link generation part. So, you have to select what are the candidate entity links and what are the all the links you have to list somehow this is called link generation part.

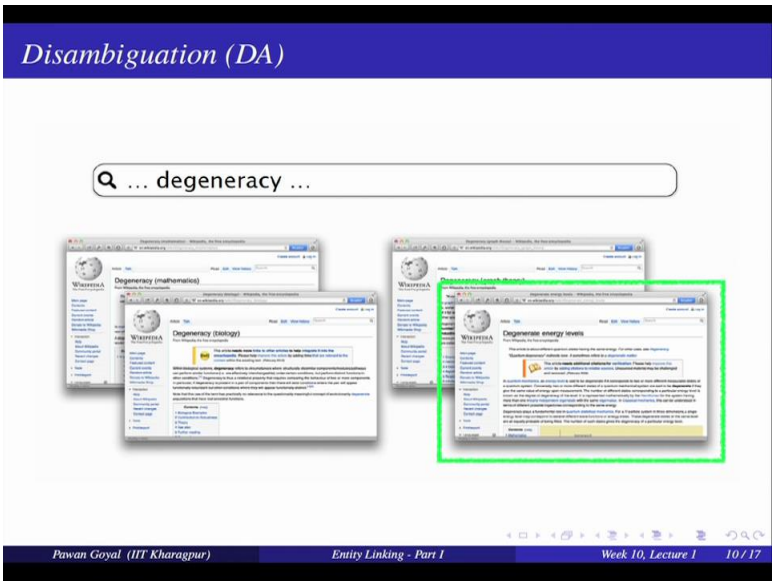
Now, once you know what are all the links then what will be the next step; you have to find out what is the most appropriate link for all these set. And this will call the disambiguation part. So, this is use the context to disambiguate what is the appropriate link it this entity should be link to and you might also want to filter you might want to improve your whole task. And we will see some examples for all these; how do you filter and improve your task based on this.

So, these are three main steps: sometimes you might combine the first two steps also, that is when you are detecting the mentions you also finding the candidates at the same time. So, that is also possible.

(Refer Slide Time: 13:03)

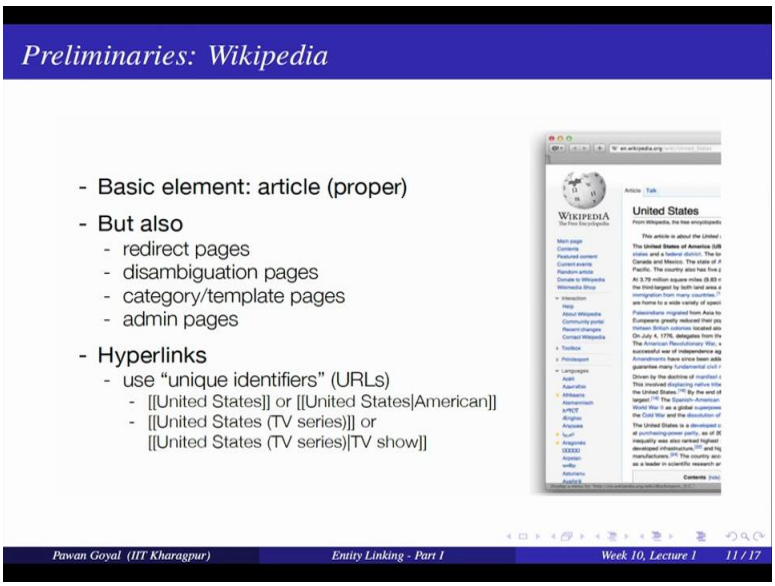
So, let us let us see these steps again in the context of Wikification. So, you are having a text where you have this sentence degeneracy is removed. And there are some words before and after. So, what is mention detection? Find out that the word degeneracy is to be linked is the appropriate mention, so that is in green in this slide. So, that is a mention detection part. Then second part would be link generation.

(Refer Slide Time: 13:25)



Find out all the appropriate, all the possible phrases pages in my database. So, here you can see the degeneracy occurs in mathematics, in biology, in graph theory and degenerate energy levels. So, there are four possible links. So, then you have a task of disambiguation. That among the four links what should be the appropriate page to which this entity should be linked, and that will be the third step and that is the disambiguation. And you will say this is the fourth one degenerate energy levels is the appropriate entity page for my mention of degeneracy. And these are three steps for this entity linking

(Refer Slide Time: 14:09)

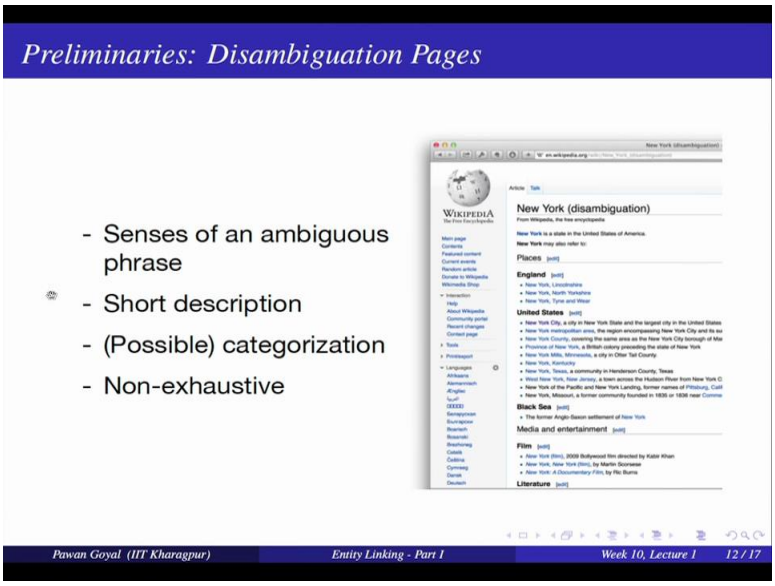


Now, so we might like to understand; what are the some of the basic features of Wikipedia that can help us in designing an algorithm for Wikification. So, Wikipedia all of you know about Wikipedia and you have been reading Wikipedia for many of your for knowing different concepts and all. So, what do you seen Wikipedia there is a page like that there is a title and certain texts about the page and you see there are various links also. So, there is an article then additionally they can be some redirect pages. So, you might have come across you are searching for something in Wikipedia and it redirects you to some other page in Wikipedia. So, these are also lot of redirect pages in Wikipedia.

Then there are disambiguation pages. We will see an example in the next slide then there are category template pages that allocate. What are the different categories in Wikipedia this category, what are the subcategories and then there are some admin pages. Now what is important for our task is that there are lots of hyperlinks in Wikipedia. So, what hyperlinks in Wikipedia? So, different words and phrases are linked within the Wikipedia itself. So, we will see that in Wikipedia article certain concepts have a hyperlink and you click on the hyperlink and you will go to the corresponding Wikipedia page. So, there are lots of in links and out links that are going on within Wikipedia.

So, United States for example; whenever you have a phrase like United States you may have a link saying it is linking to the United States TV serial or American TV show etcetera. So, you will find if you will see the source this will be like that of the hyperlink. So, you will have a double parenthesis to denote what is the appropriate entity inside the source. And that if you see the source wise you can find out. And these are various hyperlinks that you have in Wikipedia.

(Refer Slide Time: 16:23)

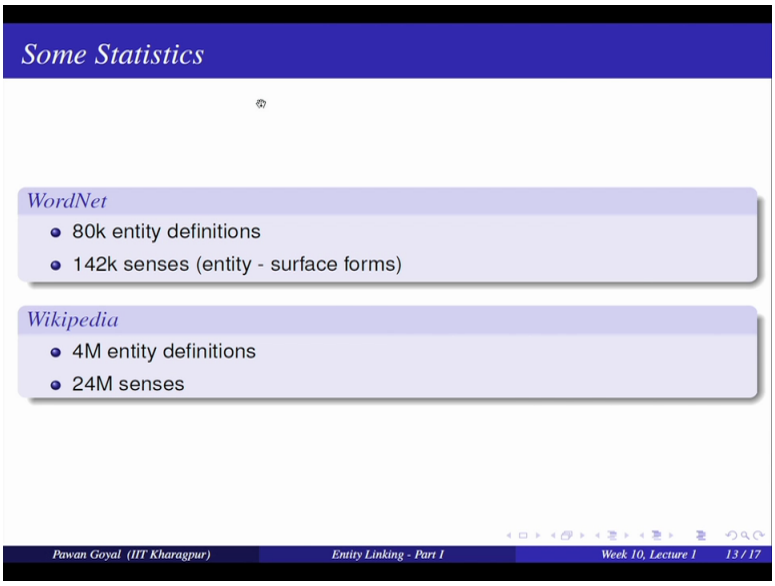


Then you have a lot of disambiguation pages. For some entities where a lot of different mentions are available you might have also categories in disambiguation; that in the category of entertainment what are the possibilities, in the category of politics what are the possibilities and so on. So, like here you have for the entity New York disambiguation page, so you will have h places what are the possibilities, then media entertain entertainment. And you will see that in each category there are lots of entities that correspond to this the main entity New York.

So, you will see in these cases is the same single entity can map to may be 15-20 different pages in Wikipedia, and they are nicely categorized in this disambiguation page. Categories may or may not be there in various pages. So, what do we need to about the whole architecture? Lot of pages in Wikipedia and each page has some name that will be the identify; you may be the identifier then lot of text involved in the text you have various hyperlinks, where different phrases are linked to their own Wikipedia pages.

And some entities will have their own disambiguation page, where you will find out what are the different ways in which this entity can be used, this maybe also under various categories.

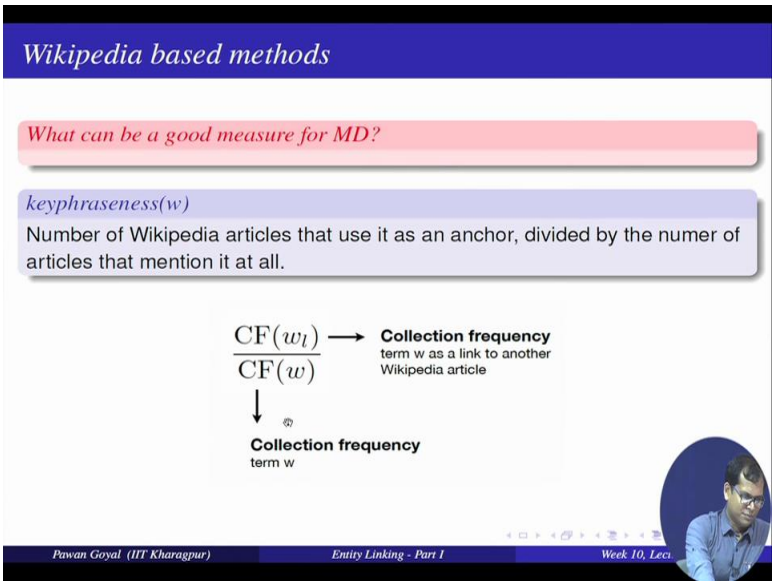
(Refer Slide Time: 17:36)



So, once you know about that let us say some small statistics. So, that is we talked about WordNet for sense disambiguation. We given a sentence, we want to find out each word; what is the appropriate sense in WordNet it corresponds to. So, in WordNet how many entities we had? We had roughly 80k; 80000 different entity definitions, and 142000 different senses.

On the other hand Wikipedia is much larger repository. So, in Wikipedia overall there are 4 Million entity definitions and this keeps on increasing, and there are 24 Million different senses. So, is much much larger in compassion to WordNet. So, our task is from all these 24 Million senses find out given a text what are the entities import that are important and which of the sense they correspond to.

(Refer Slide Time: 18:34)

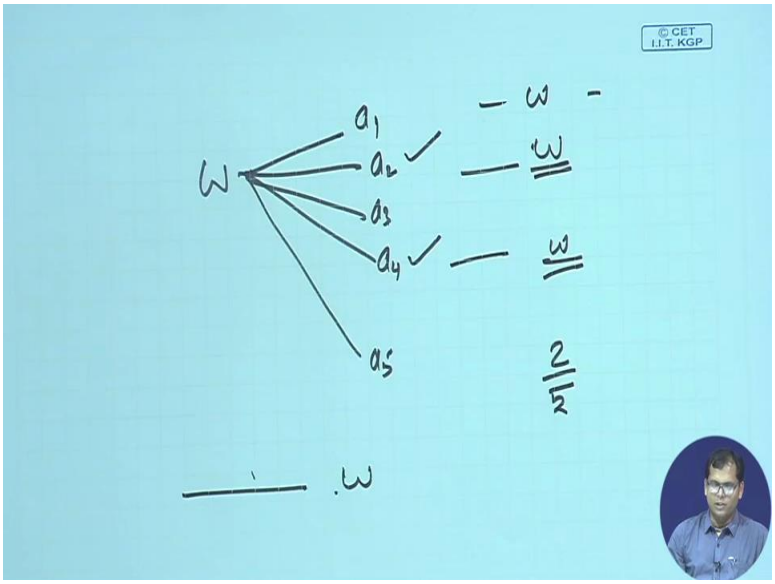


Now let us see; what are the simple measures that we can think for the three steps or let us say only the two steps: the mention detection and disambiguation. Mention detection, that is in a text whether a given n gram is in appropriate mention or not. So, what we will do initially we will see some sort of measures that can be taken simply by the Wikipedia structure or Wikipedia data. So, let us see.

So, let us talk about this mention detection part, whether a particular phrase is a good candidate for a mention. So, what will be a good measure for this? So, if you think about using Wikipedia structures we can say that- ok Wikipedia has lot of pages. Suppose I find out this particular n gram how many times it occurs in Wikipedia; and among whatever times it occur, what fraction of times it is actually linked to something. So, what is the idea if a word is linked much more number of times; that means, it might be a good candidate for mention. If it is not linked many times; that means, this may not be a very good candidate.

And this very simple criteria that you can used from Wikipedia; so in this is called the keyphraseness of a word or also a phrase. So, number of Wikipedia articles that use it as an anchor divided by the number of articles that mention it at all.

(Refer Slide Time: 20:12)

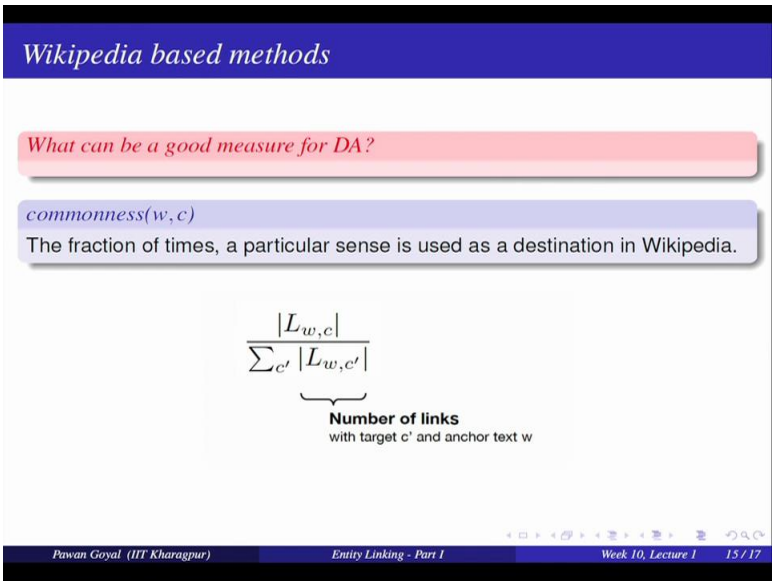


That means, I will take a word w and I will find out what are all the Wikipedia pages where it occurs. So, it occurs suppose in five (Refer Time: 20:19) articles: article 1, article 2, article 3, 4 and 5. And among the five articles say 4 and 2 provide a link with this w; so where w occurs with the link to some Wikipedia page. And a 4 also this w occurs with the link to a Wikipedia page, but a 1, a 3, a 5 do not provide a link. So, here w occurs without a link.

So, what is this keyphraseness? Keyphraseness is what fraction of the page in Wikipedia is it linked wherever it occurs. So, five linked to so keyphraseness will be 2 by 6. And that is a good measure in that it will tell me whenever encounter a new phrase w how many times it is actually linked in Wikipedia and use that to detect if it is a good mention at all. So, this is a very simple measure. So, we will say how many times it occurs and among whatever time it occurs how many times it is linked to another Wikipedia article.

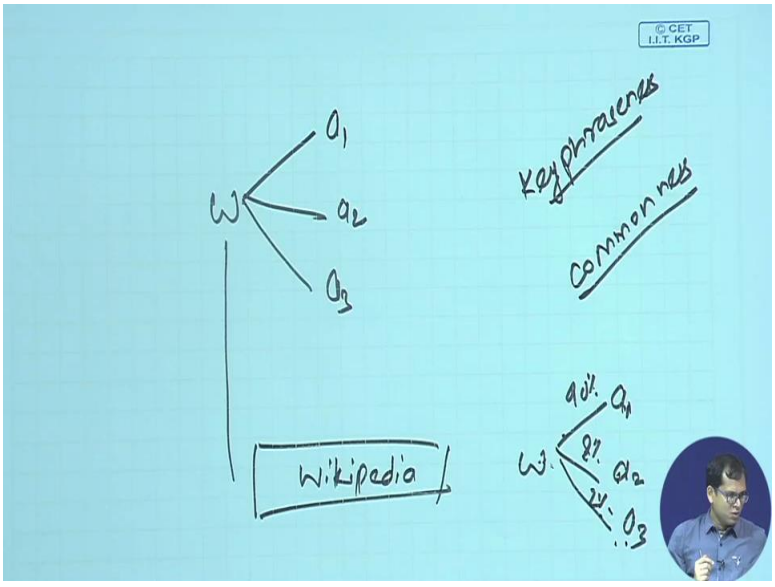
Now here we do not worry about whether it is linked always to the same Wikipedia article or multiple Wikipedia articles, the only thing is it is linked to something then we will considerate in the numerator. So, this is the keyphraseness for a word.

(Refer Slide Time: 21:41)



Now, what can be a good measure for disambiguation? So now, let us again think about it can we use Wikipedia to find out a good measure for disambiguation. So, what can be the simplest measure that you can think of? So, I have a word and it can correspond to multiple entities.

(Refer Slide Time: 22:10)



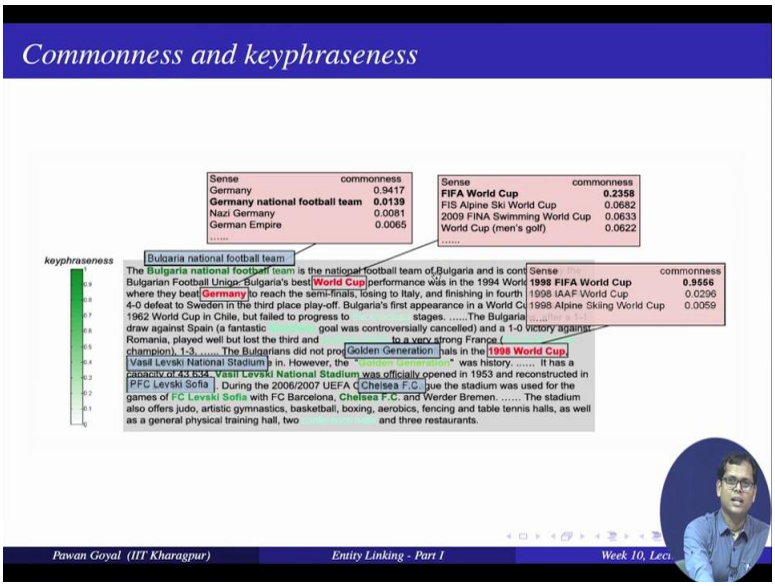
So for example; let us say I have a word w and in general it can link to three Wikipedia pages: a 1, a 2, a 3 all are possible reference for this Wikipedia page. Now, what can be a good baseline to find out what is an appropriate disambiguation page? So, for that I can

again use Wikipedia. So, I will see in the whole Wikipedia whenever w is link to something, so link to these a 1, a 2, a 3 what fraction of times it is link to a 1 suppose it link to a 1 90 percent of the time, a 2 80 percent of the time and a 3, 2 percent of the time.

And this can be a good measure to say 90 percent of times w links to the article a 1. So, by default I will say w will link to a 1. So, that can be one simple measure and this is called the Commonness; so this is commonness. So, I will define the commonness for a word and a concept; concepts here are three concepts, three Wikipedia concepts. And what is the definition? So, the fraction of times a particular sense is used as a destination in Wikipedia. So, number of times word is link to c divided by number of times word is link to any c prime

So like here it will be 90 divided by 100. Suppose they are 90, 80 and 2 pages, so 90 divided by 100 is the commonness for w and a 1 80 by 100 is commonness for w a 2 and 2 by 100 is commonness for w and a 3. So, this is another simple measure. So, you have you have seen keyphraseness and commonness. And they are simple measures it is direct from Wikipedia.

(Refer Slide Time: 24:06)



So, now let us see one example. So, here is one text that is like a report of a match, and what you are seeing? You are seeing words they are coloured and colours depend on the keyphrasenesses score; that is from 0 to 1. So, dark green is keyphraseness 1; that means,

it is always linked in Wikipedia. So, here like Bulgaria National Football team is roughly always linked to Wikipedia is has a high keyphraseness. Some words like here the knock out are not always linked they have a very low keyphraseness.

Now what about the commonness? So now you will take a particular entity like here Germany. And you will see; what are the all candidates like Germany, Germany National Football team, Nazi Germany, German Empire. Similarly for world cup it can be FIFA World Cup, FIS Alpine Skiing World Cup- 2009, FIN Swimming World Cup, World Cup Men's Golf etcetera; these are a various can candidates. And you are computing commonness by seeing how many times this word is actually linked to these entities divided by the number of times it is actually linked and this gives you the commonness.

So, Germany is linked to the Germany the word Germany like 95 percent of a time Germany National Football team 1.39 percent of the time and so on; similarly for FIFA World Cup. So now, from there you can choose by default the word the sense with the highest commonness. So, like 1998 FIFA World Cup will come up here, FIFA World Cup will come up here; but they will a problem with this entity. This Germany will written the word Germany 95 percent time, but in this case the appropriate mention is Germany National Football team and it will not be able to detect this.

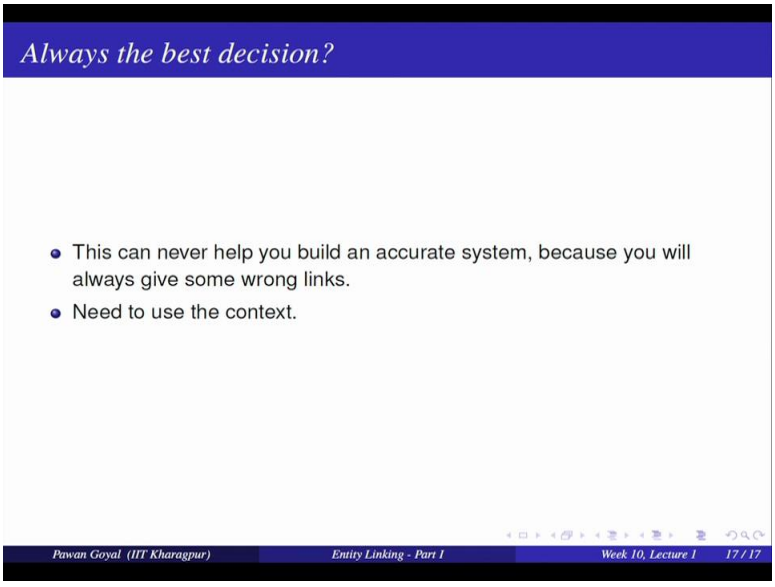
So, this is the idea about keyphraseness and commonness. And clearly you can see if there is a 1 there is one problem with this approach. So, is it always the best decision to use either the only the commonness for linking their particular entity. So what do you say from whatever we saw in the last page? Is it always the best decision to use commonness? It cannot be right, because what you are seeing whenever a word w occurs in any context I will always assign it to a 1 by default because it has the highest commonness.

That means, I have never using the context in which the word w occurs, by default I am assigning it to the category or link a 1. So that means, I will always make some mistakes right, there will be some pages at least weight should link to a 2 or a 3 and in those cases I will (Refer Time: 27:03) link it to a 1 by default. So, I cannot design a very good system by this approach. There is always the chance of making mistakes because you are taking the default case. And this also corresponds to the one of the baseline that you can use in words sense disambiguation. That is you take sense of a word that is most

probable sense. That is like a baseline, but this will never help you in designing a very good system, because you are doing it independent of the context you are always linking it to this page.

So, now what we will see in the next lecture is that can we also use the context to improve this method. Instead of using commonness can we use something from the context to find out; among the three what should be the appropriate link for this particular entity.

(Refer Slide Time: 27:57)



So, what did we see? So commonness and keyphraseness are simple measures, they can help you to design a good baseline that will work most of the times, but cannot help you build in accurate system, because you will always give some wrong links. And we can see why, because there is a default to the most probable link. And whenever the word is used in not so probable links it can never be correctly assigned.

So, we need to use the context and that is what we will see in the next lecture; that how do we use the context from the word to disambiguate the links.

Thank you.