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

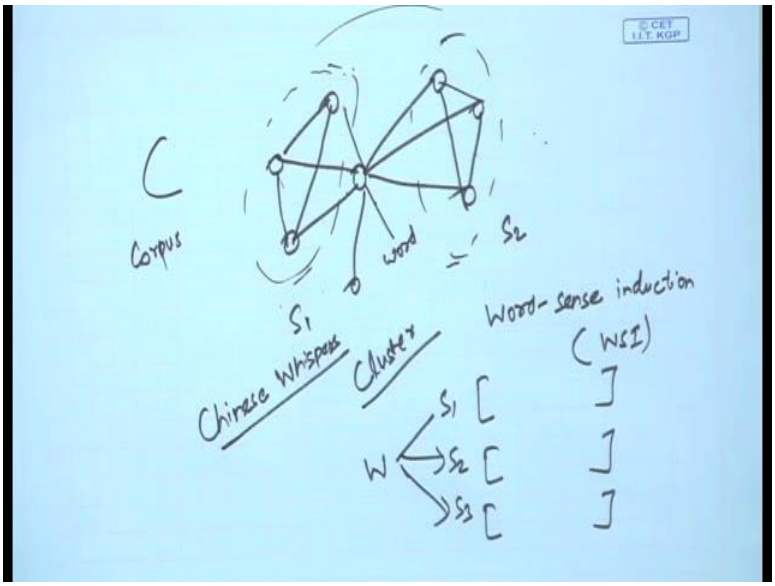
Lecture - 40
Novel Word Sense Detection

Hello everyone, so welcome to this final lecture of week-8. So, in this week, we have talked about a different method of semantics that is lexical semantics. And we saw that how we can use connection between words to find out semantics. And last two lectures, we talked about a very classical problem of word senses, disambiguation that is even a word if it has multiple senses, how do you find out in a given context what particular senses has been used. And this is a very generic kind of problem, and you can use a lot of different methods for solving that. And we saw that how you can use simple knowledge base approaches, by using the dictionary definitions and you can use some machine learning methods, you can use some bootstrapping based methods, and also unsupervised methods.

There we also talked about a page rank based algorithm if you want to jointly find out the senses for each word. We did not describe that algorithm fully. So, if you want to know in that algorithm details, so you might have to wait till we cover the summarization topic. In summarization, I will talk about page-rank in detail, but as such the generic idea you can also get from what we discussed and also you can look it up.

So, this lecture we are taking a new research problem that is coming from word sense disambiguation idea itself, and also from whatever we discussed in the towards the end of in the last lecture. So, towards the end of the last lecture, we were talking about learning the word senses in an unsupervised manner that is now I am not talking about senses defined in a dictionary. So, I am saying I have a corpus, I know how the word has been used multiple times, can I use the usage of the word to find out what are it is senses and what was the basic idea.

(Refer Slide Time: 02:23)



So, basic idea was that if I know that I have a corpus see my corpus, and I can find out which words co-occurred other words. So, like that I can construct some model I know what is co-occurred with what are others. So, idea was it suppose I am constructing this network, if a word has two senses, what will happen? The words that correspond to it is, there is suppose my center word. So, words that I connect corresponding to one sense will be connected together; words that are in second sense will be connected together. This is sense one; this is sense two. So, if I am building the whole co occurrence graph, if I center it around the particular word and try to cluster it, I can find out it is different senses. It is a very generic idea and this I call as word-sense induction also known as from WSI.

There are many ways of doing that because they are many ways you can do this clustering you can apply any graph based clustering method. So, for example, Chinese Whispers is a very popular algorithm and you can use any other methods also. So, idea would be now when you do that for each word, you will get some senses S 1, S 2, S 3 and so on. And these might correspond to say the words, these are the words, it can be different words they can have weights and so on. Now, so this is you have give a corpus, you can find out what are the word senses, but in this lecture, we will talking about the problem that can you find out if a word has got a new sense or not in some recent time, has it got a new sense. Now, does that happen? So, over the time, we keep on using the same words new, new and senses. So, and with the social media this is becoming a lot

more common that you are using the same word in some new senses that it was never used before.

(Refer Slide Time: 04:40)

So, let us see one example. So, I have the word sick right; sick is a very common and popular word and you will say what is the meaning of sick something to be related to some disease or illness. So, this is the dictionary definition, so affected with a disease or illness, of or relating to people who are ill, and very annoyed or bored by something because you have had too much of it. So, I am sick of that.

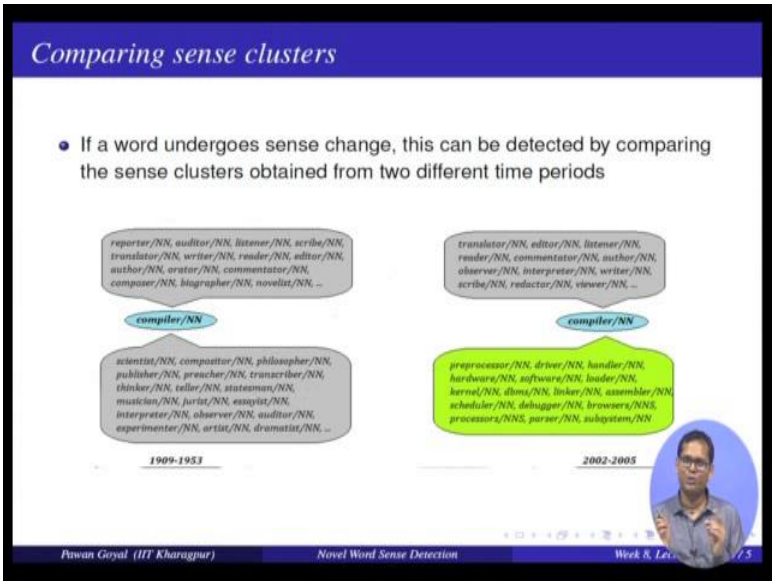
So, now this is one common sense of sick. Now, do you think sick has got any new sense in recent times, and if you think about it, sick has got a very a new sense and that is completely opposite of this particular sense. So, what is that sense? So, let us look at this tweet. So, listening to Paulo Nutini's new record, it is sick. So, now, what is the meaning of sick here of the same word sick, it is not boring. So, this sick means something that is very, very cool. So, this sick has got a very new meaning from whatever we were seeing earlier something that disease illness to something that is very cool.

Now, so what is keep on getting these new meanings in the corpus, so the way people are using that. Now, the problem is suppose my dictionary or my lexicon like word net is not getting updated regularly. So, if this word sick has got this new sense, and it is being used in this new sense in the corpus, I will never be able to match it to any of the sense in the word net, because word net has not recorded the sense at all. So, (Refer Time: 06:19)

does not know probably sick has the sense also and this is happening for many other words.

So, now what he is done in this field of novel versus rejection, from the corpus and the way the words are being used can I detect whether the word has got new sense; and if I can detect it I can populate different lexicons and different and also my word net version I can update using these definitions. So, how do we find out the word has got new sense, now again, so this is a new area, but there have been some different sort of methods and works, I will not go into details of any of those works. But what I will do? I will try to give you a basic idea that if you understand word sense induction how can you use that to find out new word sense or novel word sense.

(Refer Slide Time: 07:10)

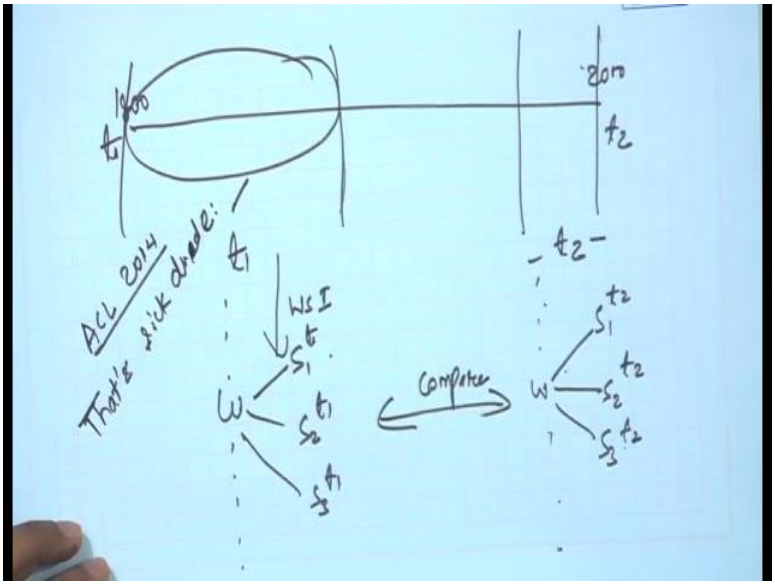


So, let us have a basic idea. The basic idea is that I can try to compare the sense clusters, so that is if your word is undergoing sense change it can be detected by comparing the sense clusters obtained from two different time periods, something like this. So, you take the word compiler, suppose in 1909-1953, so we take took the data and we did the words sense induction; and from the induction we found out that the word compiler has these two senses. So, one sense is in the sense of reporter, auditor, listener etcetera; second is on scientists, compositor, philosopher, publisher, preacher, so there are two different senses. So, what it means is that these words come together to form one sense and these

words come together to form another senses in 1909 - 1953. Compiler was merely some sort of person who wish to compile.

Now if I look at this word in recent time like 2002 – 2005, compiler has got the sense in the sense of programming language compilers, this sense was not available not there in 1909 - 1953. So, can I automatically detect that the word has got a new sense from the data itself. So, what I will do, again into (Refer Time: 08:28) I will do the word sense induction. So, suppose I do induction and I find these two senses. So, one is again the same sense translator, editor, listener, reader, commentator that is similar to what we had earlier, but you see a new sense also coming up preprocessor, driver, handler, hardware, software, loader, kernel, dbms. And you can immediately see by looking at these words in the computer sense; this sense was not available earlier so, now this is interesting observation that if I simply do word sense induction over the new time period, I find a sense cluster that was not available before and this gives me a generic framework.

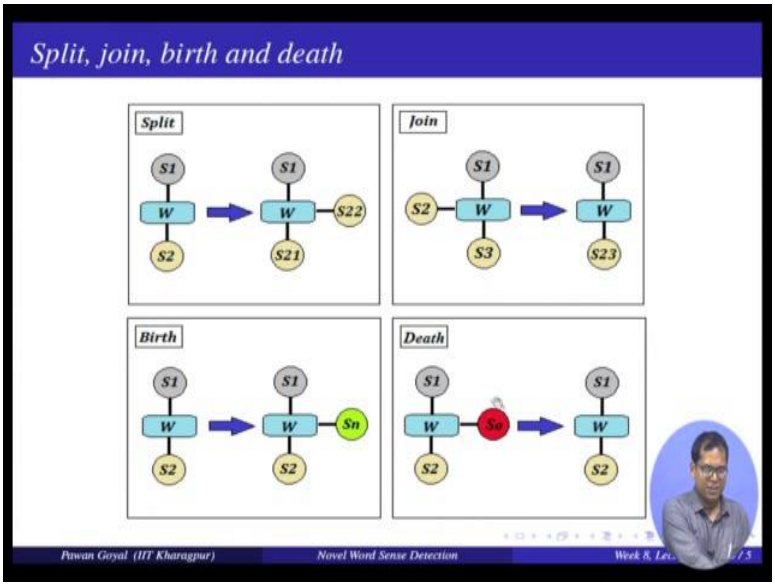
(Refer Slide Time: 09:12)



So, what I will do I will take my data that is over several time. So, it can be say 1800 to 2010 or whatever. So, it is starting from some t 1 to t 2. I want to find out what words have undergone sense change. So, what I will do, I will try to take some slice of this data call it time point t 1; take another slice later on time point t 2. So, now I will do; I will compute my co occurrence and whatever way I want to compute my distributional thesaurus then I will do word sense induction. So, for each word, I will find out s 1 in

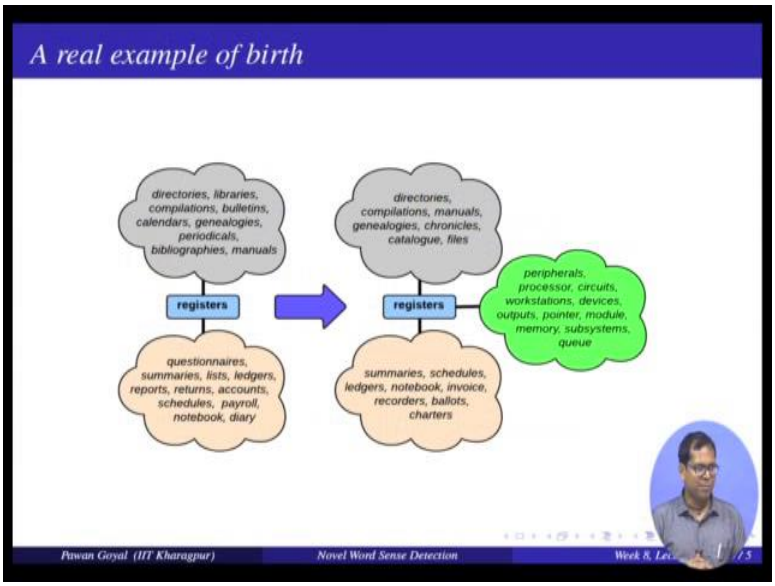
time t 1, s 2 in time t 1, s 3 in time t 1 and so on and that I will do for all the words. Similarly, I will do the same thing at time t 2; so I am at W, and I will say s 1 at time t 2, s 2 at time t 2, s 3 at time t 2 and so on. Now, I have got the sense cluster at time t 1 and at time t 1. Now, I will try to compare these to find out if the word has got a new sense cluster that was not available in the previous time point. And the second do for all the words and by doing that I can find out which words have got a new sense in the newer time period.

(Refer Slide Time: 10:42)



So, in general I can define various different sort of sense change. For example, it might happen that earlier a word has a sense, it has got to split into multiple senses or it can be a join - two different senses of joined together to form the same sense, they may not be so common. But what is common is something like a birth, that the word had initially two senses s 1 and s 2, now in the new time period, it has got a new sense s n, this is very, very common. And then they might be death of a sense also that earlier it had a sense of s 0, now in the new time period I cannot sense find out the sense. So, all these you can find out just by comparing these sense clusters.

(Refer Slide Time: 11:29)



So, here is another example that we observed from the data. So, this is from the work that we did in 2014. So, this was in ACL 2014, if you want to have a look. So, the title was "That's sick dude" and there was some other subtitle, so this was in ACL 2014. So, we took the data from Google anagrams, and we took the whole data from starting from somewhere around 1600 to 2008 divided into eight different time points then so that they were divided such that in each time point you have roughly the same amount of data.

So, as you go over recent years even like three years and four years for making complete time duration; in earlier it might be 100 years together because the data was not too much. And then we took different time points, constructed distribution thesaurus and then you Chinese (Refer Time: 12:34) algorithm to find out sense clusters and then complain the clusters. So, what are the words that are getting some new sense? So, here was an one example. So, we found that the word registers earlier had these senses, and you can see these are like dictionaries, directories, libraries, compilations, bulletins these are registers as you like paper registers and so on so that you have even now; similarly, here notebook, diary, returns, accounts. So, in the new time period, we found that these two senses are there directories, compilations, manuals, summaries, schedules, ledgers, notebook, but very new senses come up. And can you think of the sense what is the sense peripherals, processor, circuits, workstations devices. So, register has got the sense of if the sense of computing in a hardware that.

So, this new sense we could detect simply by using the corpus that these all words for being used with resistance and they were having similar co occurrence such as the word to resistance. So, we found this is a new sense cluster that is coming up and like that we did in many other senses. So, this is as I was saying this is a new research field and some works have come up in this field this is one such work, but I hope the basic ideas care that if you want to detect new word sense what you need to do, so that ends this week of lexical semantics.

So, next week again we will continue with the topic of semantics. So, we have started with distributions of semantics - one idea, then we talked about lexical semantics another idea of semantics. Now, we will see how we can capture semantics using topics, can we find out what are the topics that are there in my data use that for some semantics and that is why we will discuss topic models in detail. And there are again very, very popular tools in NLP.

So, thank you, I will see you in the next week.