Natural Language Processing

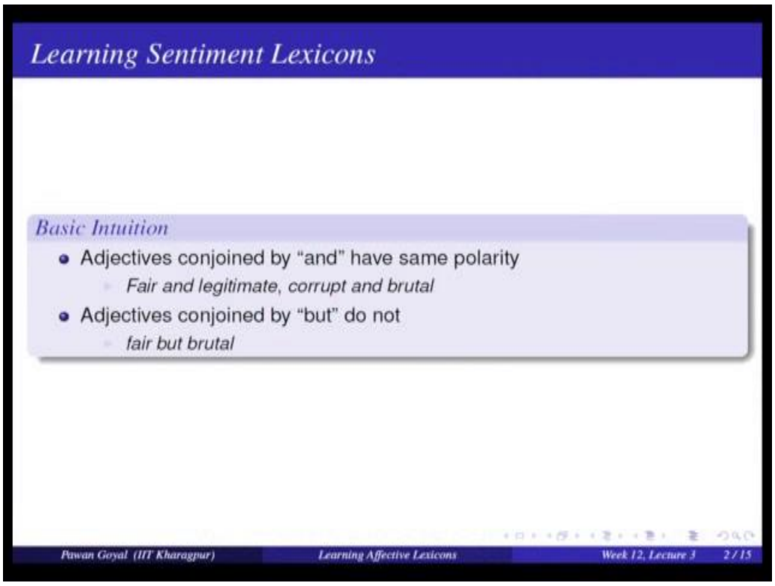
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 58
Learning Affective Lexicons

Hello everyone. Welcome to the third lecture of this week. So, we were talking about sentiment analysis. And in the last lecture, we talked about various sentiment lexicons that are available to us and that you can use for various task related to sentiment analysis. So, in this lecture, what we will be doing we will be seen suppose you have to learn these lexicons on your own without having to do manually labeling each word with the sentiment polarity, so what can be some possible approaches you can take. So, while you will see that most of the lexicons are built for English language they not too many works for other languages.

So, suppose you want to build a lexicon for your own language, it can be Indian language, for example, so you can want to build for Hindi, Bengali and or Tamil, so how do you start approaching this problem. So, you can always take all the words and have people manually annotate that, but is there some automated approach and that is why you will also see some of the concepts that we have talked in this course how they can be useful in doing this task.

(Refer Slide Time: 01:19)



Learning Sentiment Lexicons

Basic Intuition

- Adjectives conjoined by “and” have same polarity
  - *Fair and legitimate, corrupt and brutal*
- Adjectives conjoined by “but” do not
  - *fair but brutal*

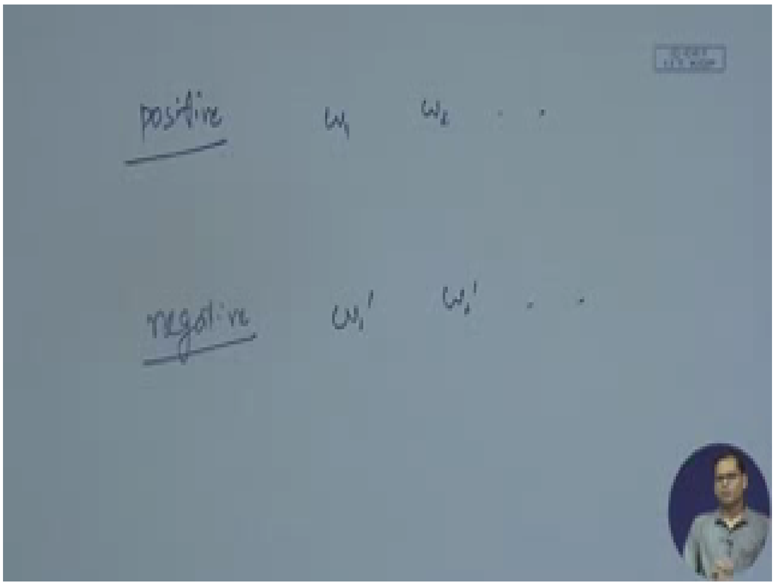
Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 12, Lecture 3 2 / 15

So, when suppose I have to built, I have to learn these the polarity of different words in my lexicon. So, what might be the basic idea intuition that I can start with. And then intuition can come again from some sort of distribution hypothesis in that when to the words of say simple idea get to together or how do the words of different polarity are get together. So, for example, here we are saying think of the adjectives that are joined by and. So, will they have the same polarity?

Suppose, you are having a sentence and you see fair and legitimate. So, you can say probably both these words have same polarity. You may not be able to tell; what is the polarity of the individual word, but if they occurring with and fair and legitimate, you would be able to see that they will have the same polarity. Same here, corrupt and brutal these two words have the same polarity. Now, can you have some similar indication for saying whether the two words will have different polarity? So, think about cases where you say, but and so on like here it is fair, but brutal. So, you will know that the polarity of both the words fair and brutal will be opposite to each other.

Now can this sort of simple intuition can help us in learning the sentiment lexicons. So, what will you have to do if you have to use some methods like that? So, we can see that we cannot find out probably what is the polarity of an individual word by this method, but if I know the polarity of a single word, I might be able to infer the same for other words if they occurring with some sort of connectives like and, or, but. So, what should I start with?

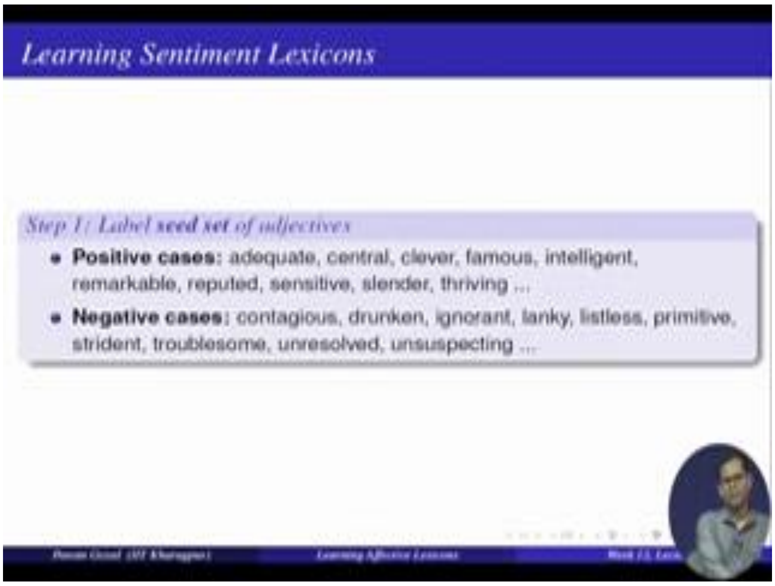
(Refer Slide Time: 03:04)



So, what I might be doing I might be starting with list of positive words and some list of negative words positive and negative sentiment words. So, I take some w 1, w 2 so on w 1 prime w 2 prime and so on. So, this might be some very simple list that you start with that you might have created manually. So, what you will be doing next, you know there the words that I have in the same polarity might occur in a corpus with certain connectives.

So, now you will go to a large corpus, it can be web or any of the corpus that you have and try to find out these if these words are occurring with other words with some connectives. And for that you can just search like fair and, you can search for this and wherever you find fair and x you might assign a positive polarity to x. Similarly, fair, but and you find fair, but brutal and you might assign a negative polarity to that so that can even possible approach.

(Refer Slide Time: 04:00)



Learning Sentiment Lexicons

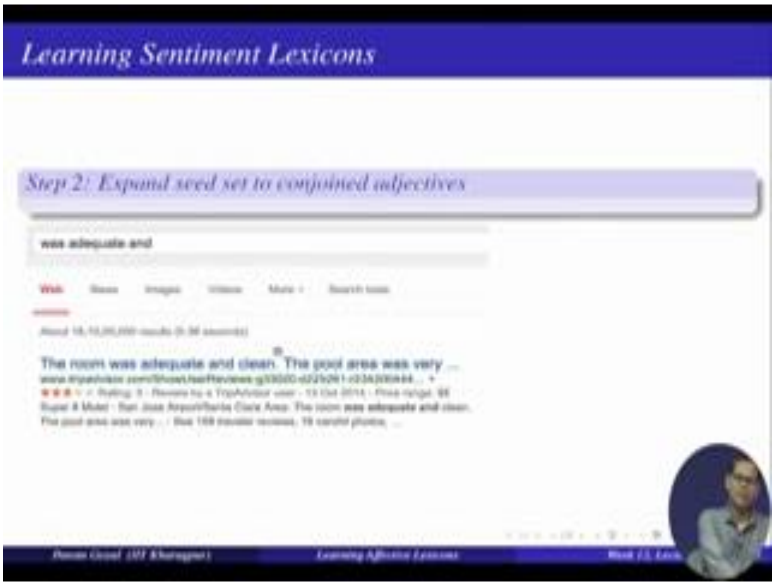
Step 1: Label ***seed set*** *of adjectives*

- **Positive cases:** adequate, central, clever, famous, intelligent, remarkable, reputed, sensitive, slender, thriving ...
- **Negative cases:** contagious, drunken, ignorant, lanky, listless, primitive, strident, troublesome, unresolved, unsuspecting ...

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 11, Lec

So, I start with some seed set of adjectives that I can label manually or I can get it from some sort of it can be obtained also from some websites give me some positive and negative adjectives. Now, so suppose I got some positive cases adequate, central, clever, famous; and negative cases like contagious, drunken, ignorant, and so on. So, now I go to my corpus and try to search these with some connectives added to these.

(Refer Slide Time: 04:30)



Learning Sentiment Lexicons

Step 2: Expand seed set to conjoined adjectives

was adequate and

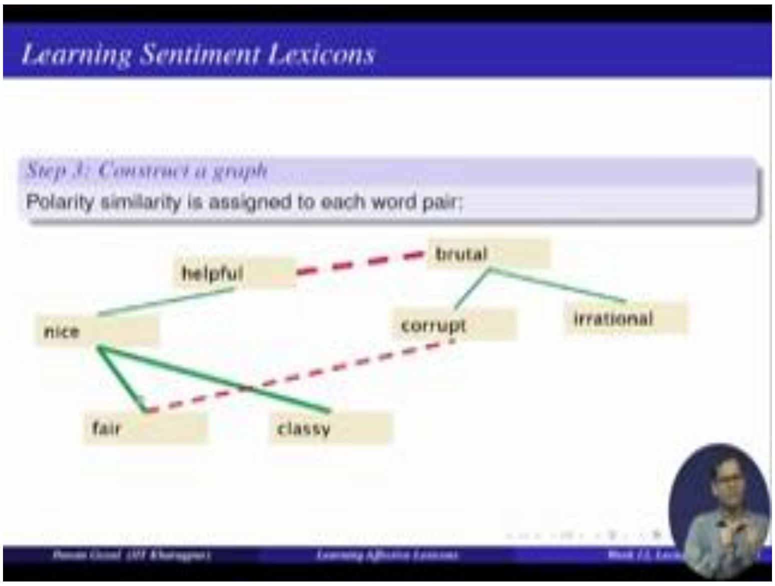
About 18,10,00,000 results (0.36 seconds)

The room was adequate and clean. The pool area was very ...
www.tripadvisor.com/ShowUserReviews-g33020-d225261-r234200444...
★★★☆☆ Rating: 3 - Review by a TripAdvisor user - 13 Oct 2014 - Price range: $$
Super 8 Motel - San Jose Airport/Santa Clara Area: The room **was adequate and** clean. The pool area was very ... - See 159 traveler reviews, 19 candid photos, ...

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 11, Lec

So, like here, so I want to expand my seed set to conjoined adjectives. So, I can search say on web. So, I will say was adequate and, so that way I will find out the words that are having a possible ideas similar to adequate. So, suppose I search that I find some results like this. The room was adequate and clean and immediately, I will say clean will probably have a positive polarity tool.

(Refer Slide Time: 04:57)



Learning Sentiment Lexicons

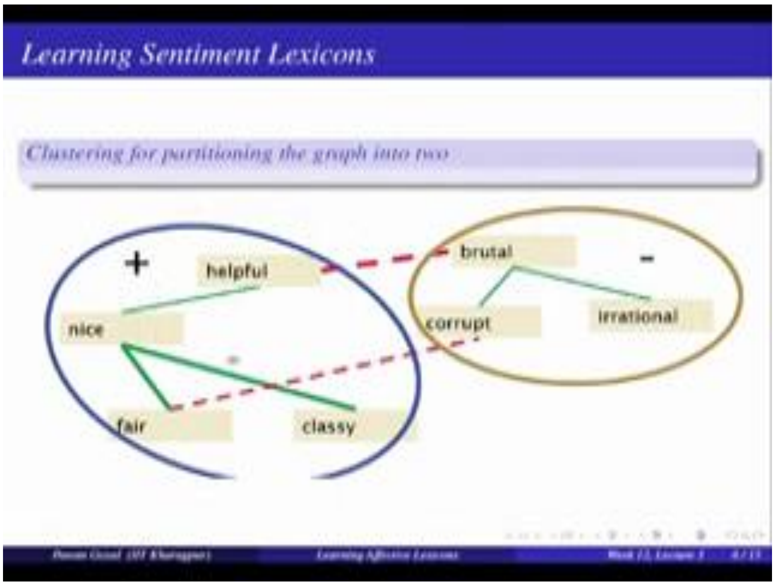
Step 3: Construct a graph

Polarity similarity is assigned to each word pair:

And if I do that if I apply this method over all my seed sets, also from the seed set I get other adjectives, you can continue applying this method and you can obtain a graph like that. So, what I using in this graph you find that nice and fair occur with some connectives, so that is why there is an edge green edge and they are up appearing with the connective like and that means, same polarity. Similarly, here nice and helpful occurring with connective end, nice and classy occurring with connective end, fair classy do not occur together, but that may not be required.

Similarly, you find brutal and corrupt, they occur when this with the connective, end brutal and irrational occur with the connective end, but there are some red dashed lines also, what are these so that would be suppose towards are occurring with the connective of, but so helpful, but brutal. So, then you say they have opposed polarity. So, now, that is how you will label the edges in your graph. So, some are edges that see ok they have similar pole polarity some edges that is why they have polarity and now once you have built this graph over a large set of words you can apply some clustering algorithm. To say which of these words are having green edges with among each other, but no red edges that can be some criteria by which you can cluster these words. And this can finally, give you positive and negative examples from the data.

(Refer Slide Time: 06:30)



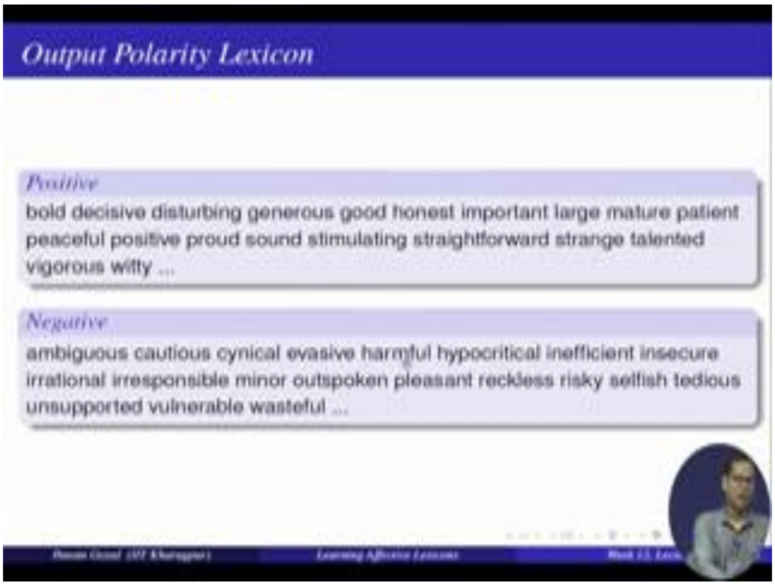
Learning Sentiment Lexicons

Clustering for partitioning the graph into two

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 12, Lecture 3 6/15

So, that is the partition of the two graphs and different graphs give you positive as well as negative sentiment lexicons that can be the standard method that to be applied over any language once you have some seed set of sentiment words and you have a corpus at your hand. And you can always increase your set of rules that you are having so right now you are using only and or, but you find out some other rules by which different sentiment words can be connected, and you apply those rules also to get different edges in your graph. So, once you apply this method from a real experiment what kind of lexicon did they obtain.

(Refer Slide Time: 07:13)



Output Polarity Lexicon

Positive

bold decisive disturbing generous good honest important large mature patient peaceful positive proud sound stimulating straightforward strange talented vigorous witty ...

Negative

ambiguous cautious cynical evasive harmful hypocritical inefficient insecure irrational irresponsible minor outspoken pleasant reckless risky selfish tedious unsupported vulnerable wasteful ...

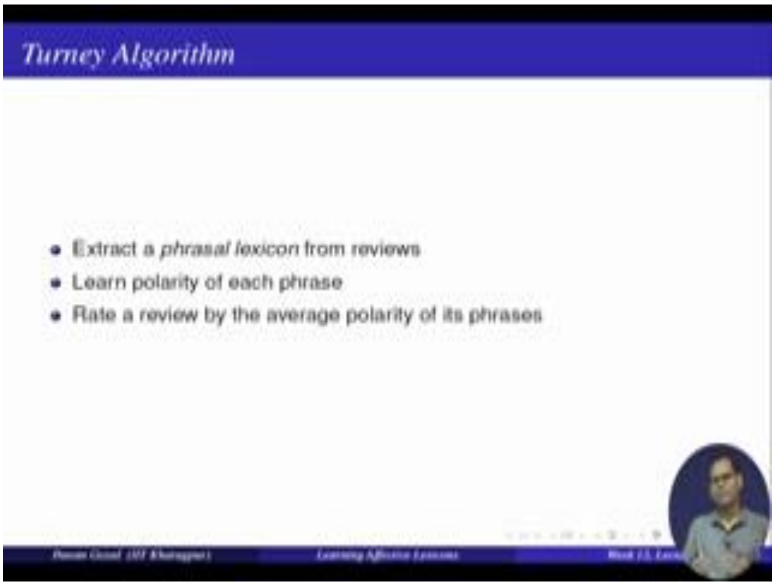
Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 11, Lec

You see the positive words the obtained words like bold this is disturbing, generous, good, honest, important, large, mature and so on. These are the words that they obtain obtained this will be noisy this may not be very, very good set, but the ideas is that you will get a lot of good examples right. So, is like generous, good, honest, important, large, mature, peaceful, they look all having they all look like having a positive sentiment. The only exception looks like disturbing here they may be some other little.

Similarly, that these are the words that they obtain for negative sentiment ambiguous, cautions, cynical, evasive, harmful, hypocritical, and all these look the words with negative sentiment. So, this was obtained by giving a seed set, seed label to only to a few words, not these words. And these when you found out the words with different connectives, you are able to obtain this positive and negative class. And there will be some erroneous cases like disturbing, coming as positive example, cautions; outspoken pleasant coming as negative example that few cases might be there, but overall this get good precision.

And if you have a very good corpus, good seed set you might also get a very good (Refer Time: 08:29) by this method. So, that is one simple method that you can apply to some sort of bootstraps your sentimental lexicon.

(Refer Slide Time: 08:39)



Turney Algorithm

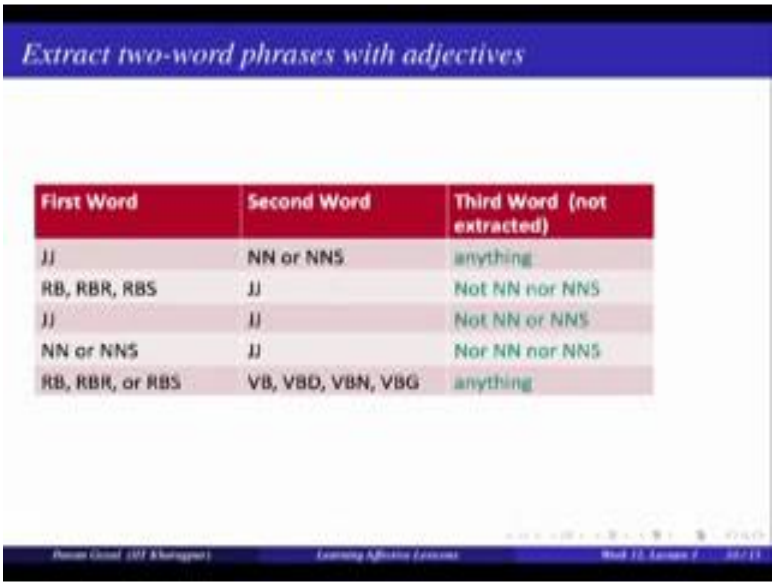
- Extract a *phrasal lexicon* from reviews
- Learn polarity of each phrase
- Rate a review by the average polarity of its phrases

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 11, Lec

Now so there are some other algorithms also propose a literature for doing this task. So, we will see some of these. So, let us look at the Turney algorithm. So, what did this algorithm? So, what this algorithm did. So, they took some example from reviews. So, they start with the data first, they did not start with seed set of labeled words with emotions opinions, they start with the data set. And they said because you got a lot of opinions from review data set. So, let us take some review data sets. Now, in the review data set they are so what the hypothesis was so opinions will be expressed by some sort of noun phrases, there are some adjectives they are some nouns and there will be opinions associate with those.

So, let us start with extracting these noun phrases that are occurring a lot in this corpus. Once we have a good set of noun phrases then I will try to find out some sort of sentiment associate with those. So, let us see how they did that. So, you first extract a phrasal lexicon from reviews, and then try to learn polarity for each phrase here. Once you have learned the polarity, you might go to a next task that is find out what will be the rating of this review. Suppose, this is a plane text there is no rating, you can use the sentiment from faces in the review to give the rating to the review, whether it is a positive review or negative review.

(Refer Slide Time: 10:06)



Extract two-word phrases with adjectives

First Word	Second Word	Third Word (not extracted)
JJ	NN or NNS	anything
RB, RBR, RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Nor NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN, VBG	anything

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 11, Lecture [illegible]

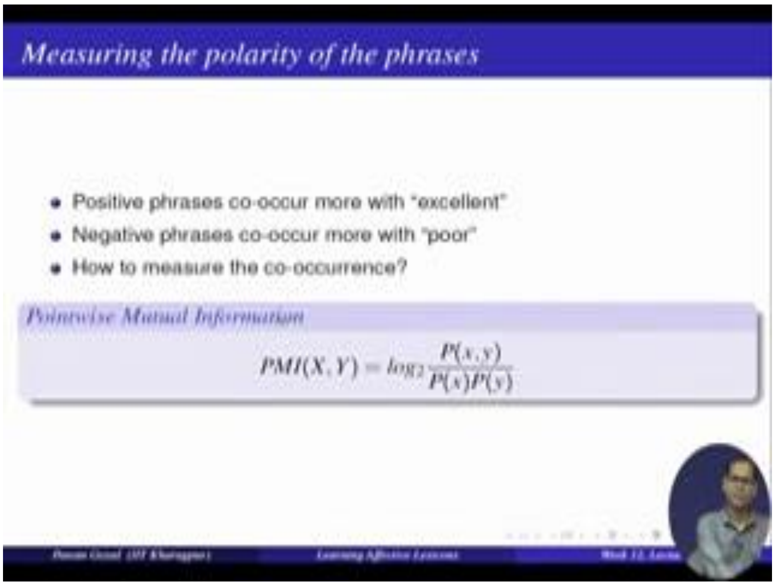
So, interesting thing was how to start building their phrasal lexicon. So, it did not they did not take any noun phrase that occurs in the review. So, they made some patterns by manually seeing how the important sentiment opinion phrases occur in text. So, what kinds of patters are there? And they built some patterns like the first word should be a JJ an adjective second word can be a NN or NNS. And third word not extracted is anything that is whenever they find first word as an adjective part of speech text second word with NN or NNS, they will extracted if there free the lexicon irrespective of what is the third word they will do with all the corpus. Like they built other rules like if it is in adverb, second word is an adjective, and third word is not noun or NNS, they will take it in their phrasal lexicon.

Similarly, both first and second word are adjectives third word is not NN or NNS, they will take it. First word is NN or NNS, second word is adjective and then the next word is not NN or NNS again they will take it in their phrasal lexicon and like that they had another rule adjective and a verb. So, whatever phrases in my lexicon or the conjugative words my lexicon were following these regular expressions defined to a part of speech text they took that in their phrasal lexicon. Again that is a very interesting method you can take some examples on your own and find out what is the usual pattern of part of speech sector that occurs and according you define that is how you will extract your phrasal lexicon.

So, now once they extract their phrasal lexicon, the next task would be how do they give them some sentiment score. So, what the hypothesis they were saying some of these will positive some of these be negative. So, whichever are positive will probably occur with a word like excellent; and whichever negative will occur with words like poor. So, now how do we measure this co occurrence? So, how many times it is occurring with excellent how many times it is occurring with poor and use, that to give a sentiment polarity to each of these word in the lexicon.

So, again here you can use the corpus. So, you can take a large corpus find out how many times each of these phrases occurring with excellence and poor, but simple count will not work. So, we have to use a very sophisticated method. So, do you remember any such measure that we discussed in this course. So, to find out how common do they do towards co occur. So, if you remember we talked about point wise mutual information, so you can find out the point wise mutual information of each word with these excellent and poor.

(Refer Slide Time: 13:02)



Measuring the polarity of the phrases

- Positive phrases co-occur more with "excellent"
- Negative phrases co-occur more with "poor"
- How to measure the co-occurrence?

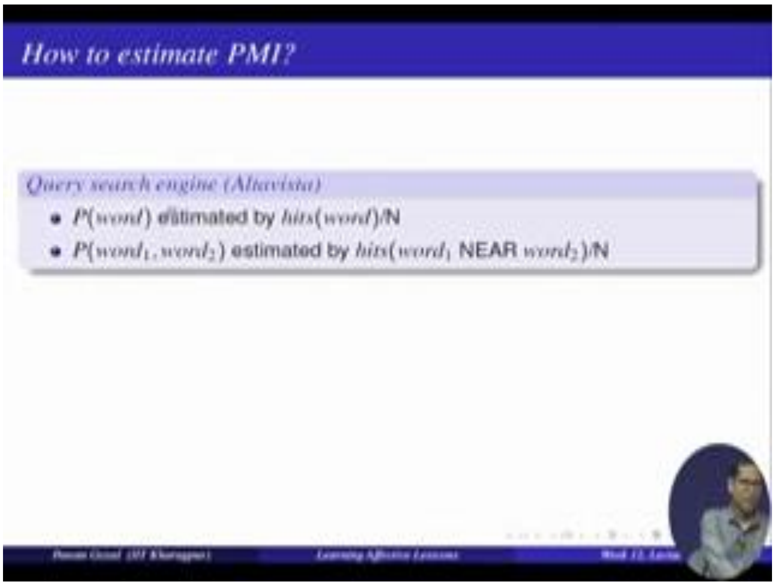
Pointwise Mutual Information

$$PMI(X,Y) = log_2 \frac{P(x,y)}{P(x)P(y)}$$

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 12, Lectu

So, I will use this method to find out what is the PMI between a phrasal lexicon and excellent phrasal word and excellent and phrasal word and poor. So, how did they find this PMI value, you are always use a corpus for this, but they also did something interesting. So, at that time, the main search engine was AltaVista. So, what they did they tried these queries over AltaVista. So, they said ok.

(Refer Slide Time: 13:31)



How to estimate PMI?

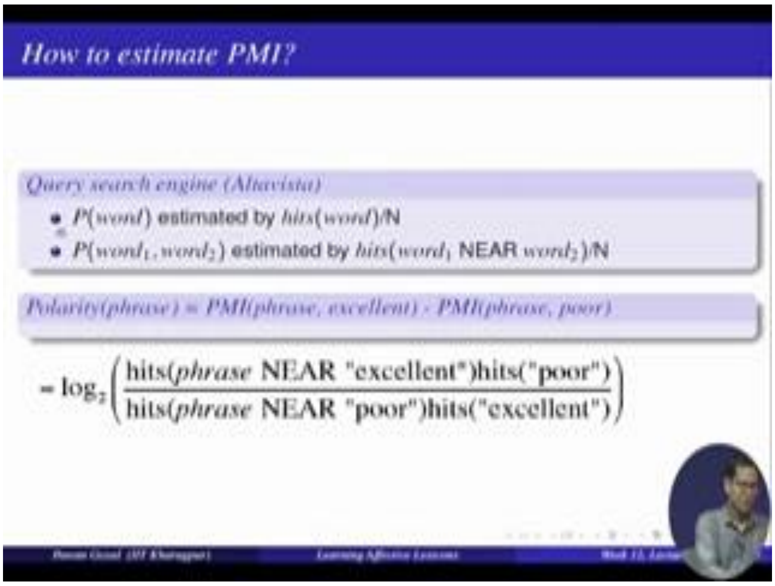
Query search engine (Altavista)

- $P(word)$ estimated by $hits(word)$/N
- $P(word_1, word_2)$ estimated by $hits(word_1$ NEAR $word_2)$/N

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 12, Lectu

So, they have the estimate the three things probability of word probability of excellent, and probability of word and excellent occurring together. And they were using how many hits the word or the two words together are getting by the AltaVista engine that is again a nice method of computing the probabilities and co occurrence probability of this word. So, you give that a query find out how many hits the search engine is giving you. So, I have to assume a probability word, I see how many hit the word get by the engine divide by N. Probability word from word to again how many hits word 1 near word 2, some query like that get over the engine divide by N. Now, once I have these two measures, I can easily compute the PMI score.

(Refer Slide Time: 14:16)



How to estimate PMI?

Query search engine (Altavista)

- $P(word)$ estimated by $hits(word)/N$
- $P(word_1, word_2)$ estimated by $hits(word_1 \text{ NEAR } word_2)/N$

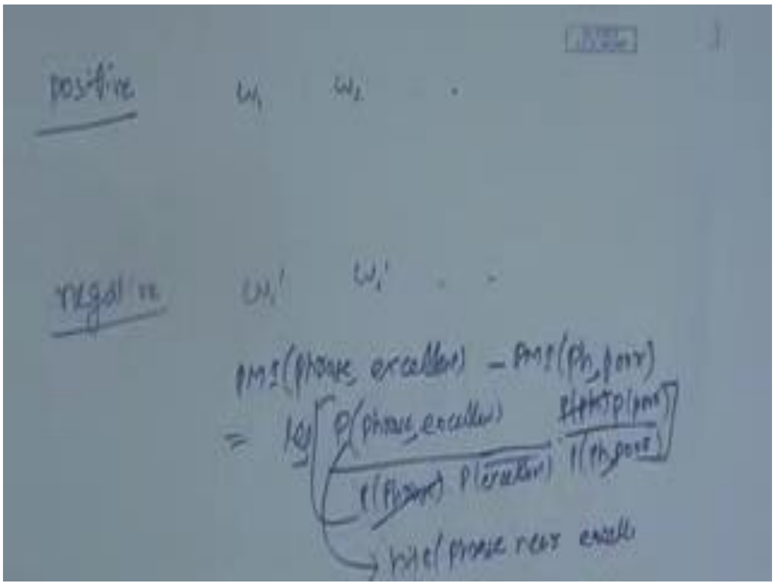
Polarity(phrase) = PMI(phrase, excellent) - PMI(phrase, poor)

$$= \log_2\left(\frac{\text{hits}(\textit{phrase} \text{ NEAR "excellent"})\text{hits}(\text{"poor"})}{\text{hits}(\textit{phrase} \text{ NEAR "poor"})\text{hits}(\text{"excellent"})}\right)$$

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 11, Lecture

So, what is the polarity of the phrase that is PMI of the phrase with excellent minus PMI of the phrase with poor and that will give you this particular expression. If you want we can just quickly do that on paper.

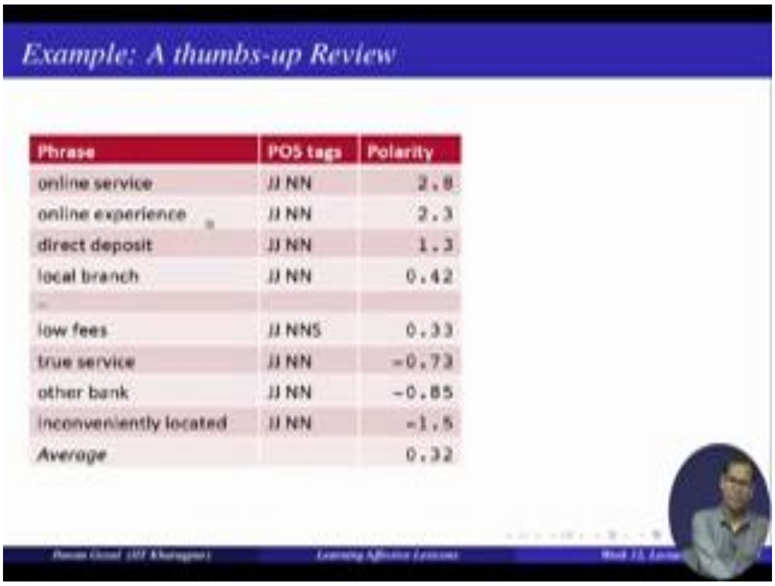
(Refer Slide Time: 14:35)



positive
w_1 w_2 .
negative
w_1' w_1' . .
PMI(phrase, excellent) − PMI(Ph, poor)
$= \log\left[\frac{P(\text{phrase, excellent})}{P(\text{phrase})\,P(\text{excellent})} \cdot \frac{P(\text{ph})\,P(\text{poor})}{P(\text{ph, poor})}\right]$
hits(phrase near excell

So, we have to get PMI phrase excellent minus PMI phrase poor. So, how do I get that is equal to log probability phrase excellent divide by probability phrase probability excellent is. Minus log becomes you have to I men you are reversing it. So, it will become probability phrase poor probability phrase probability poor that will come out to be the PMI. And then you can immediately say that this is repeating. So, you have only this one log probability phrase excellent, and this you obtain using hits phrase near excellent. Similarly, phrase near poor, this is hits for excellent, hits for poor and that is how you get this polarity of the phrase. And you do that for all the phrases that you have taken in your opinion lexicon you obtain some sort of polarity for this. So, these phrases occur with excellent a lot, these occur with poor a lot.

(Refer Slide Time: 16:00)



Example: A thumbs-up Review

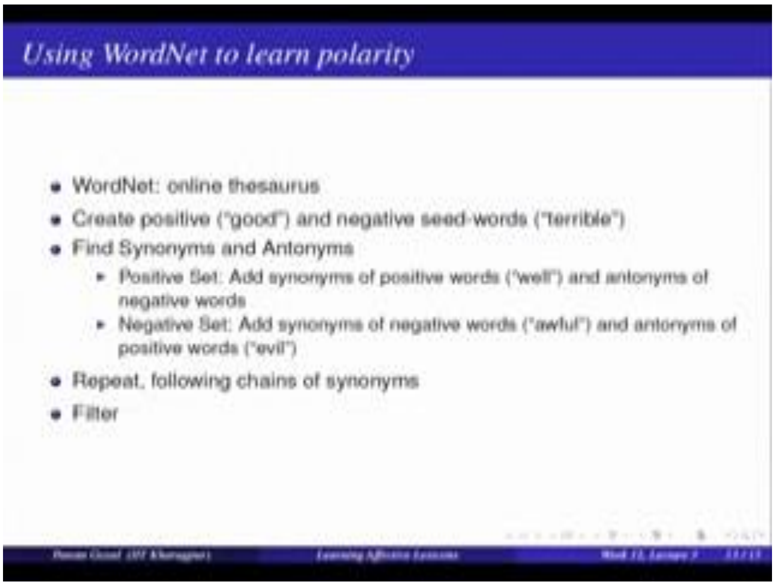
Phrase	POS tags	Polarity
online service	JJ NN	2.8
online experience	JJ NN	2.3
direct deposit	JJ NN	1.3
local branch	JJ NN	0.42
...		
low fees	JJ NNS	0.33
true service	JJ NN	-0.73
other bank	JJ NN	-0.85
inconveniently located	JJ NN	-1.5
Average		0.32

Pawan Goyal (IIT Kharagpur) Learning Affective Lexicons Week 11, Lecture

So, if you see from thumbs-up review, what are difference is that are they obtain, they obtained words like online service with this part of speech tags JJ NN polarity 2.8, online experience JJ NN 2.3, direct deposit JJ NN 1.3. And then they found out what is the average polarity score of each phrase that they got and they said something like 0.32. You can see there were some negative examples also there like low, true service but probably not negative, but inconveniently located first negative and they got a minus 1.5. So, they got some positive negative examples even in thumps up review, but the average that they obtained was positive 0.32.

Similarly they did for negative the thumbs-down reviews from the product site and while they were again something like direct deposits, online web, very handy with positive polarity they were many with negative polarity like virtual monopoly, lesser evil, other problems, low funds, unethical practices and they were you can see you have very, very negative score by their method and that was very nice. So, that was very interesting approach. So, by that they were able to obtain such phrases and also tag them with the polarity values.

(Refer Slide Time: 17:00)



Using WordNet to learn polarity

- WordNet: online thesaurus
- Create positive ("good") and negative seed-words ("terrible")
- Find Synonyms and Antonyms
  - Positive Set: Add synonyms of positive words ("well") and antonyms of negative words
  - Negative Set: Add synonyms of negative words ("awful") and antonyms of positive words ("evil")
- Repeat, following chains of synonyms
- Filter

So, these are some algorithms there are many other algorithms literature, but one other method that you can always use is to use your WordNet. So, all of you know WordNet we talked about that in one of the earlier lectures. So, how do you use WordNet. So, again the idea is some sort of bootstrapping; in the WordNet find out some positive and negative sentiment words. Suppose, you find out this word is positive this word is negative. So, how do you bootstrap from there, how do you find more words? So, one simple thing that you can think is that find out all the synonyms of this word. So, all the synonyms would also can be given a positive polarity, if it I positive; if it is negative, they can be given a negative polarity this is one way.

Then you can find out the antonyms and give them the negative polarity; if it is positive it should be antonym should be given negative polarity; if it is negative and antonyms should given positive polarity. And once you get more seed set, you can continue building over there. So, this can be one approach using WordNet. So, you create some positive that is good and negative seed words. So, positive is like good and negative like terrible. So, once you have the seed set you find the synonyms and antonyms.

So, wherever whatever are the synonyms of the of the positive word like good, you add to the positive set; and whatever the antonyms of the of the negative word like terrible you again add to the positive set, and you do the reverse for the negative set. So, add synonyms of negative words and antonyms of positive word. So, like you will get well in the positive set, and awful and evil in the negative set.

And this you can further repeat now, you have got more words, you can keep on repeating them. And you might create some new set seed sets whenever you feel this is not giving me for that examples you might create some more seed sets and this might be again some simple nice method of build starting to bootstrapping your opinion lexicons, you can also combine different approaches together. So, you start some words from WordNet go to the other approach the first approach that we talked about using and, and but you can also start from their and then go to WordNet and to find an more synonyms, so both are possible.

And again there are many more methods, but I think this will be some interesting ideas that you can use for building your opinion lexicon. So, this was about this lecture. In the next lecture, what we will do we will see how you can also obtain the also see some nice trends about different words being used in the reviews, say review with rating 1 to 10 how are the different words, the positive and negative words are being used. And what are the nice ways in which you can use the sentiment lexicons that we have talked about. So, I will see you in the next lecture.

Thank you.