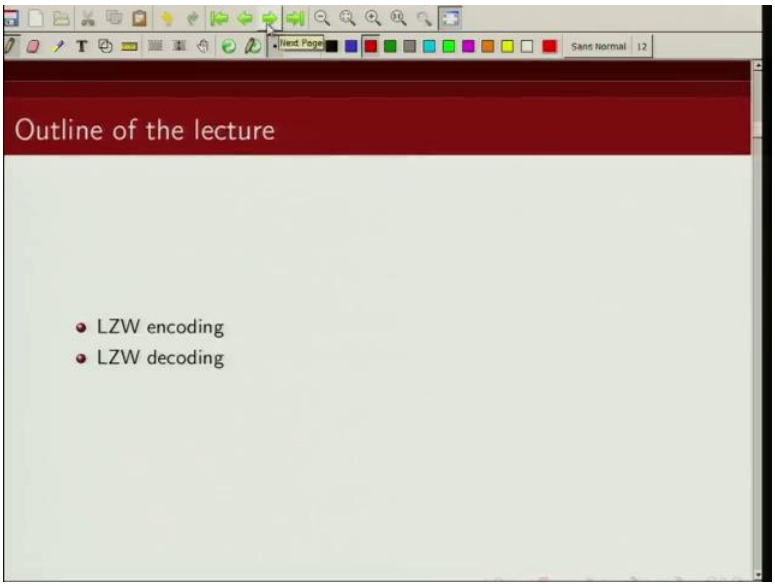
An Introduction to Information Theory
Prof. Adrish Banerjee
Department of Electronics and Communication Engineering
Indian Institute of Technology, Kanpur

Lecture - 07B
Universal Source Coding-II: Lempel-Ziv Welch Algorithm (LZW)

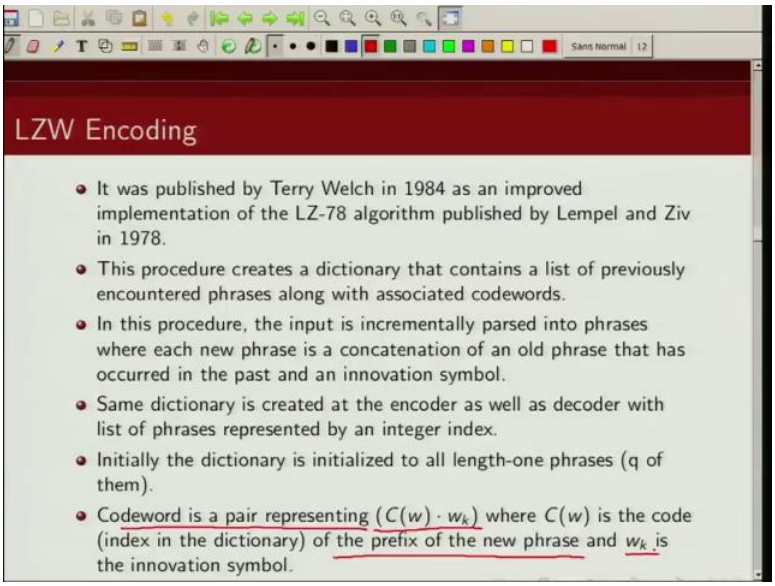
Welcome to the course on An Introduction to Information Theory. In today's lecture, we are going to continue our discussion on Lempel-Ziv algorithm. Today, we are going to talk about another variant of Lempel-Ziv algorithm, which is Lempel-Ziv Welch algorithm.

(Refer Slide Time: 00:37)



So, we will in today's lecture, we will talk about Lempel-Ziv Welch encoding as well as decoding and we will illustrate this with an example.

(Refer Slide Time: 00:46)

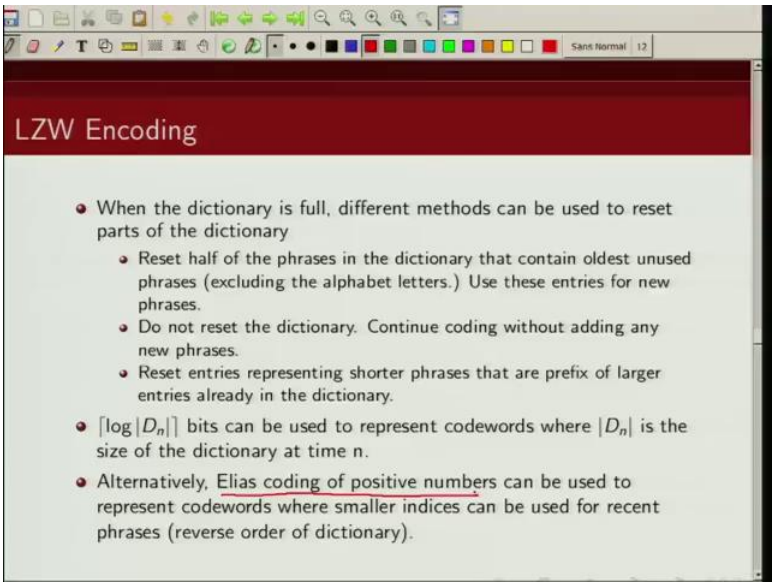


So, in 1984, Terry Welch published an improved implementation of Lempel-Ziv 78 algorithm which was published by Lempel-Ziv in 1978. And we are going to talk about this algorithm. So, in this algorithm, we create a dictionary that contains a list of previously encountered phrases along with a associative code word. So, we are going to create a dictionary of previously encountered phrases, and we will put them in the dictionary. In this procedure, the input is incrementally parsed into phrases where each new phrase is a concatenation of an old phrase that has already occurred in the past and a new symbol, which we will call as innovation symbol.

So, we are going to look for a sequence which is already there in the dictionary plus a new symbol which is innovation symbol and we will put that in our dictionary. So, as I said in this method, we are incrementally parsing the input into phrases where each new phrase is a concatenation of an old phrase that has already occurred in the past and it is there in dictionary and a new symbol which we are calling an innovation symbol. And the same dictionary is created at both the encoder as well as the decoder. And we list of these phrases in dictionary is index by a integer index.

To start with, we will initialize our dictionary to all one-length phrases. So, let say your output is binary. So, we will put 0 and 1 in our initial dictionary. And a codeword is pair representing the code that is index in dictionary and of the prefix of the new phrase and w k is the innovation symbol. We will talk about this when we explain it within example.

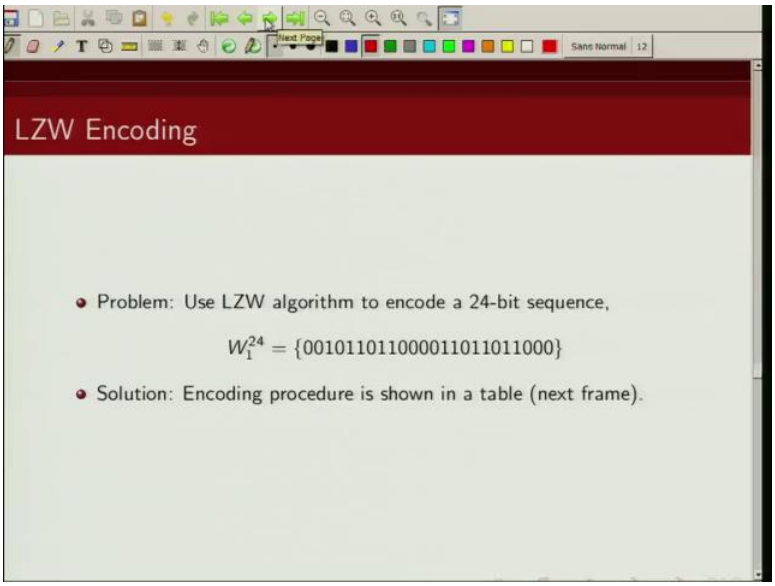
(Refer Slide Time: 03:19)



So, as we noticed in the previous slide what we have are doing here is we are adding new phrases into our dictionary which is already not there. Now, if our dictionary gets full because our dictionary typically will have fixed size then what should we do. So, we can follow one of these two approaches. In this approach, we can reset half of the phrases in dictionary that contain the oldest unused phrases. So, phrases which have been their dictionary, but we cannot appearing in the new phrases like phrases which have not been used. Those old phrases can be reset that we create space for entering new phrases or the second approach could be we do not reset the dictionary we continue coding with the same set of phrases that have already there in dictionary. So, we do not update the dictionary and we continue coding using the older phrases. Another approach could be we reset entries corresponding to short phrases that are already prefix of larger entries and these larger entries are already there in dictionary.

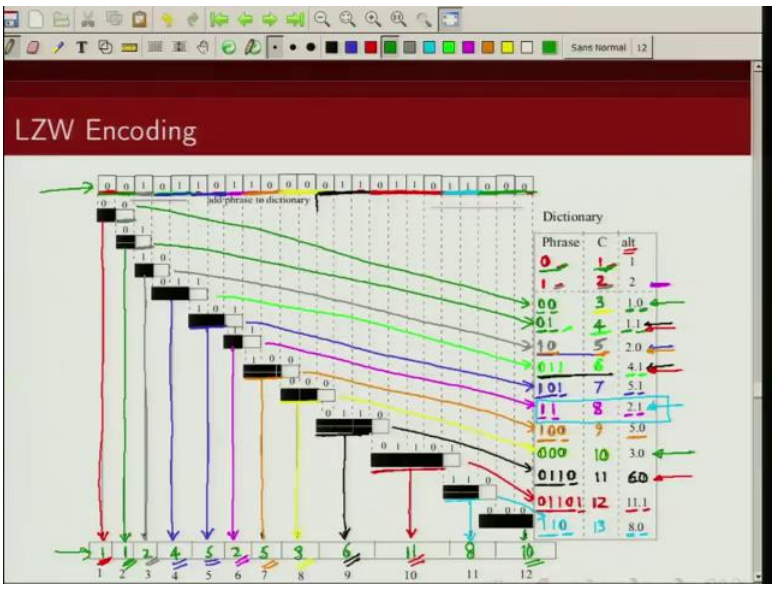
So, for example, if you encounter 0, 1 and if 0 1 1 0 or 0 1 1 1 or 0 1 0 0 0 1 0 1 those are already there in dictionary then that probably not much point keeping 0 1 so that is what I am meant when I said reset entries representing short phrases. So, this is the size of the dictionary, so we will require these many numbers of bits to represent our codewords. Ultimately, we can also use a coding of positive numbers, which was given by Elias we can use that to encode these code words. Now, this Elias coding of positive number, we will talk about in the next weeks lecture.

(Refer Slide Time: 05:45)



So, let us take an example to illustrate how Lempel-Ziv Welch algorithm encoding will work. So, this is 24 bit binary sequences that evolve to encode using Lempel-Ziv Welch algorithm. Now, remember this is a universal source comprehension algorithm. So, this algorithm does not need pair information about the distribution of zeros and ones.

(Refer Slide Time: 06:18)



So, this is the sequence that we want to encode that is 0 0 1 0 1 1 0 1 1 0 0 0 0 1 1 0 1 1 0 1 1 0 0 0 this is the sequence that we want to encode using Lempel-Ziv Welch algorithm. So, what is the first step, the first step is to initialize the dictionary with all length 1

phrases here we are talking about binary sequence. So, those length one phrases are 0 and 1. Here, interesting them by 1 and 2. Alternative representation, I talk about in a little while let me first complete this table. So, the 2 length was sequence is 0 which indexed by code word 1; and this phrase 1 which is index by code word 2. So, initially my dictionary will contain 0 and 1.

Now let us look at the encoded the sequence I want to encode. So, the first the bit here is 0, but that is there in the dictionary to zeroes. What is our next bit 0 0. So, this next to the 0, if 0 0 in there in dictionary, no, the only bit in the dictionary is this phrase 0 and phrase 1. So, we can only encode 0 there and what is the encoding of 0 0 corresponds to this code word 1, so that is why you see 0 must be encoded using 1 here.

Next we looked at the next bit that is 0. Now, if 0 0 in the dictionary at this point 0 0 is not there in dictionary. So, what we will do we will put 0 0 in the dictionary. So, we will add 0 0 in the dictionary and that is a codeword number 3. Now, this alternate representation as I said each new phrase can be written as concatenation of an old phrase and new innovation symbol. So, this 0 0 can be written as 0 concatenated with a new symbol 0. And what is 0, 0 is my code word 1. So, this codeword 3 can be written as this code word index by 1 and this new innovation symbol which is 0. So, the ultimate representation of this is the phrase divided by codeword 1 and then followed by a innovation symbol 0.

Now, so we put 0 0 in dictionary. So, now, let us look at this 0 here, the next bit is what next bit is 1; 0 1 is not there. In the dictionary, you saw I can only encode 0. Please note I am trying to find the largest phrase which is already there in the dictionary and that I have been encoding at. So, see 0 1 is not there in the dictionary then only encode 0 and what is 0, 0 is my codeword 1 as you can see here 0 is my codeword 1 is my code word. So, 0 is my codeword 1. So, this will be coded as 1.

Next, we are looking at 1 here, let us look at the next bit that is 0, is 1 0 in the dictionary no. Now, remember here when we encountered 0 followed by 1. So, this was not there in the dictionary. So, what we have to do, we have to put 0 1 in the dictionary. So, I put 0 1 in the dictionary that is make codeword 4. From here one can written as codeword corresponding to 0 concatenated with a in the ratio symbol which is 1. And what is the codeword for 0 corresponds to codeword 1. So, this can be written as code word 1

concatenated with a new innovative symbol 1. So, we are at this point here 1, now 1 0 is not there in the dictionary. So, we will only going to encode 1. And what is 1 one corresponds to a code word 2. So, this will be codeword 2. So, so far we have encoded 0 0 1 that is codeword 1 and codeword 2.

Now as we can see looked beyond 1, 1 0 is not there in the dictionary. So, what are we going to do we are going to put 1 0 in the dictionary. So, 1 0 we put it in the dictionary and that is my code word index by 5. Now, we continue our encoding. So, the next bit is 0. We have looked beyond this 0 1. If 0 1 in the dictionary, yes, it is there in the dictionary, what about 0 1 1 is it there in dictionary no, it is not there in dictionary. So, what are we going to do, we are going to put 0 1 1 in the dictionary. So, we are going to put 0 1 1 in the dictionary, and that is my codeword number 6. And note that this can be written as code phrase 0 1 concatenated with a innovative symbol 1. And what is 0 1 0 1 is my codeword four. So, you can write in alternate representation I can write this as 4 concatenated with the innovation symbol 1.

So, 0 1 1 I put in the dictionary. Now, I am going to encode this. So, it is going to 0 1, 0 1 is 4. So, you note this will be 0 1 by corresponds to 4. Next, start with this 1, this 1 there in dictionary very much sure that is this one. What about 1 0, is 1 0 in the dictionary yes, it is there. Now, let us look at 1 0 1; if 1 0 1 in the dictionary not so far. So, what we are going to do is we are going to put 1 0 1 in the dictionary. So, we will put 1 0 1 in the dictionary and that is indexed by this codeword number 7, and we can write it as code corresponds to 1 0 concatenated with innovation symbol 1. So, it is 1 0 1 0 is codeword 5, and this 1 is the innovation symbol. So, you can see that each new phrase that I am adding in my dictionary can be written as an old phrase concatenated with the new innovation symbol. So, 1 0 or 1 I have put in the dictionary, and now I am going to encode 1 0. So, what is 1 0, 1 0 corresponds to 5. So, then this will corresponds to with 5.

Next, start encoding 1, is 1 in the dictionary, yes, 1 is here. 1 1, is 1 1 in the dictionary, the answer is no, we have 0 1 0 0 0 1 1 0 0 1 1 1 0 1, but 1 1 is not there in the dictionary. So, what will we do, we are going to put 1 1 into a dictionary so that is one one that is code word indexed by 8. And this can be written as concatenation of 1 with this innovation symbol 1 and this 1 corresponds to code word 2 concatenated with innovation symbol 1. So, we are going to code 1 then 1 corresponds to codeword number 2. So, we

start from here this 1, 1 is already there in the dictionary; 1 0, is 1 0 in the dictionary, yes, there this 1 1 0 is there in dictionary.

Let us look at the 1 0 0, is 1 0 0 there in dictionary not so far. So, then we are going to put 1 0 0 into the dictionary so that will be 1 0 0, this is code number 9. This can be concatenation of 1 0 with innovation symbol 0, and 1 0 is code word 5. So, it is 5 concatenated with innovation symbol 0. So, now that we have put 1 0 0 in the dictionary, we are going to encode 1 0 and 1 0 corresponds to code word 5, so that is why I have it 5 here. Next, I start with this 0, now this 0 already in the dictionary, 0 0 is 0 0 in the dictionary, yes. So, it continued further 0 0 0, is 0 0 is there 0 in the dictionary the answer is no. So, I am going to put this 0 0 0 in the dictionary that is code word number 10. And I am going to encode 0 0, 0 0 corresponds to code word number 3. So, this is 3 here.

Next, we start with this one, because we have already encoded up to this point. So, this is 0, 0 is already there in dictionary. Start with 0 1, is 0 1 in dictionary, yes, it is that is this one. What about 0 1 1 if 0 1 1 in the dictionary the answer is yes that is this. I look at 0 1 1 0, is 0 1 1 0 in dictionary and we can see it is not there in the dictionary. So, we are going to put 0 1 1 0 into the dictionary. So, we are going to add 0 1 1 0 into dictionary and that is codeword 11. This can be written as concatenation of codeword corresponding to 0 1 1 with this new this new innovation symbol is 0.

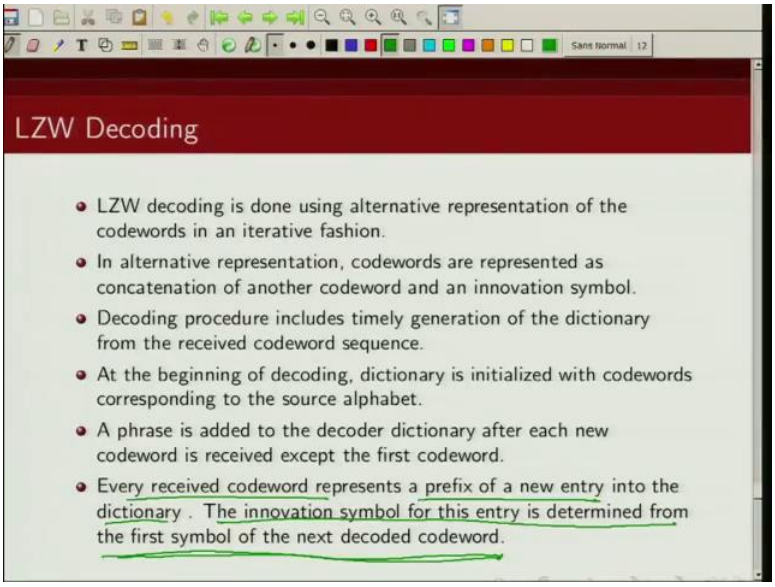
What is 0 1 1 0 1 1 is this codeword 6, 6 concatenated with innovation symbol 0. Now, we are going to encode 0 1 1, which is codeword number 6. Now, we start from this 0 zero is already there in dictionary 0 1 0 1 is also there in dictionary that is this 0 1 1 that is also there in dictionary, 0 1 1 0 this is also there in dictionary, then 0 1 1 0 1 this is not there in the dictionary. So, what we are going to do is we are going to put 0 1 1 0 1 we are going to put this in the dictionary. So, we will put 0 1 1 0 1 in the dictionary and that is a codeword number 12. And this can be written as concatenation of this codeword corresponding to 1 to 0 1 1 0 which is codeword number 11 and this innovation symbol 1.

Now, we encode this sequence 0 1 1 0 which is codeword number 11 so that is code word number 11. And we start from this 1, 1 is already there in dictionary 1 1, 1 1 is also there in dictionary that is this one. 1 1 0, is 1 1 0 is in the dictionary the answer is no we

do not see any 1 1 0. So, we are going to add 1 1 0 into dictionary. So, this we are going to add into dictionary, and these are codeword number 13 this can be written as concatenation of code word corresponding to 1 1 which is this code number 8 and concatenated with an innovation symbol that is 0.

Now, we are going to encode this 1 1, 1 1 is encoded as 8 that is this 1. So, now, we start encoding from this point to 1 1, we have already encoded. So, start with 0, 0 is already there in dictionary; 0 0, 0 0 is also in there in dictionary that is this 1 0 0 0 is it there in dictionary, the answer is yes that is this 1 cod word number 10; 0 0 0 is this 1. So, then 0 0 0 can be encoded as 10. So, I were Lempel-Ziv encoding of the sequence which is given here is basically coding of these numbers 1 1 2 4 5 2 5 3 6 11 8 and 10. So, this is how we are building up the dictionary for Lempel-Ziv Welch algorithm, and this is how they are encoding our sequence.

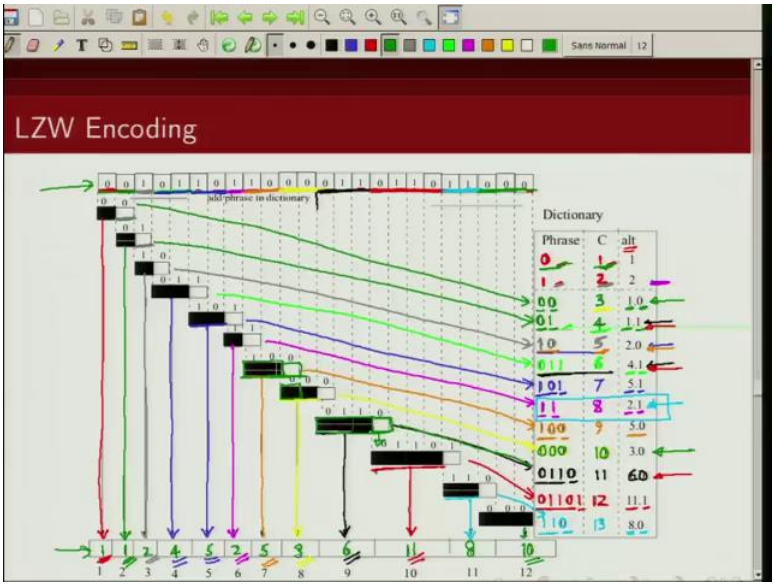
(Refer Slide Time: 21:57)



Now, let us move to the decoding of Lempel-Ziv Welch algorithm. So, LZW decoding is done using an alternative of the code word in an iterative fashion. So, as I said in the alternative representation, the codewords are represented as concatenation of another codeword and innovation symbol. So, decoding procedure includes the timely generation of a dictionary from the receive sequence and we will illustrate that with the help of an example. So, at the beginning of decoding, we are going to initialize the dictionary using codewords of length 1.

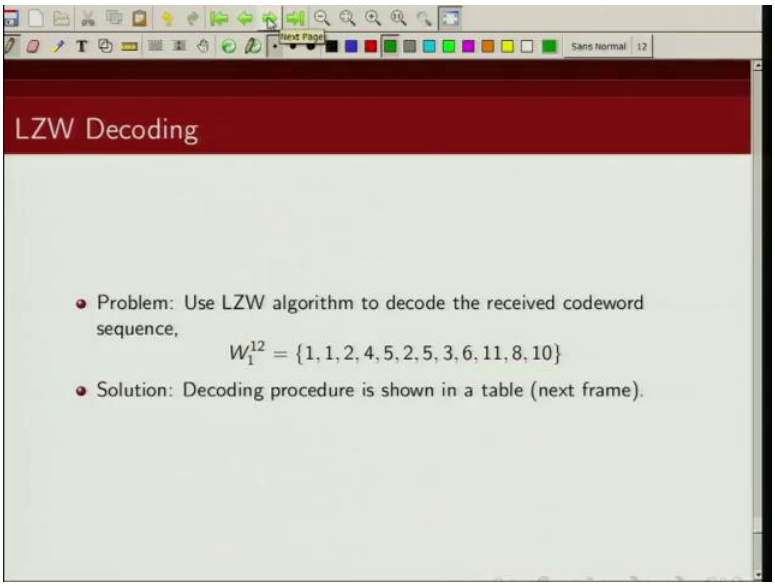
So, in the previous example, we what considering binary sequence. So, we will initialize the dictionary at the decoded with 0 and 1 in this example. A phrase is added to the decoder dictionary after each new code word is received except the first codeword. So, you will see that we are going to add new phrases into a dictionary after we decode the sequences. And every received code word represents a prefix of a new entry into the dictionary. The innovation symbol for this entry is determined from the first symbol of the next decoded codeword. If you call when we were encoding using LZW, so we were putting a string a new string into a dictionary which is already not there which we are concatenation of some string which is already there plus innovation symbol. So, this new innovation symbol was getting encoded in the next phrase, so that is why the innovation symbols or an entry can be determined from the first symbol of the next decoded codeword.

(Refer Slide Time: 24:06)



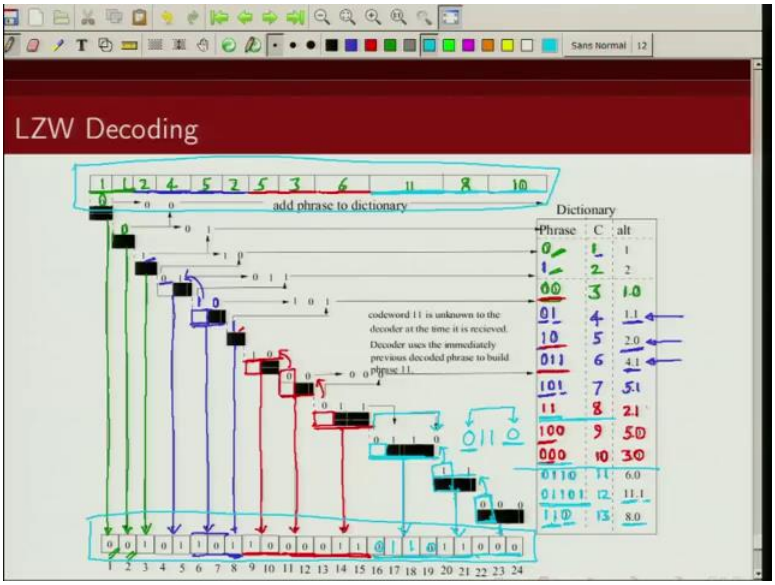
If you go back here and look at this example at in each case when we are let say we are encoding this. So, what we are doing is we are encoding 1 0 and this new innovation symbol get gets encoded in the next sequence same here. We were putting 0 1 0 into the dictionary that the new innovation symbol here was 0 this was getting coded in the next set of sequence. So, this was getting coded here. You can see basically each of these innovation symbols we are getting coded in the next phrase and that are why I made this remark that the innovation symbols for this entry is determined from the first symbol of the next decoded codeword.

(Refer Slide Time: 24:59)



Now, let us take an example to illustrate that. So, this is the same sequence we just coded it and this was the sequence in which this sequence was coded. So, let us try to decode that our original sequence. So, Lempel-Ziv coding will give us basically coding of these positive numbers. So, on these positive numbers, we have to get back our original sequence.

(Refer Slide Time: 25:33)



So, let us look at the decoding. So, what we have received is this 1 1 2 4 5 2 5 3 6 11 8 and 10 this is what we have received. On the first step in decoding is to initialize the

dictionary, and how are we going to initialize the dictionary, we are going to initialize the dictionary with length 1 phrases. We are talking about binary sequence here. So, those length wise sequences will be 0 and 1. So, 0 is my codeword indexed by 1, 1 is my codeword indexed by 2. Now, next I start the decoding what I received is 1 or what is 1 codeword 1 corresponds to phrase 0. So, this will be decoded as 0. So, this will be decoded as 0. Next, I decode this, the next encoded sequence are receive as 1, what is 1, 1 again correspondence to phrase 0. So, this will corresponds to phrase 0. So, this is phrase 0.

Now, what do I need to do the first bit of this new decoded phrase is my innovation symbol. So, I need to add to this the innovation symbol. So, I am going to add 0 0 So, I am going to add 0 0, why 0 0 because earlier force was 0 innovation symbol here is 0 So, I am going to add 0 0 to my dictionary. So, I add 0 0 into dictionary that is code word number 3. And in the alternative of representing with form this is 0 concatenations with 0. So, 0 is codeword 1 concatenation with 0.

Next, I receive 2. So, 2 correspond to bit 1 so then this will be decoded as bit 1. And what is the next thing I need to do, I need to add 0 1 into my dictionary. So, the old phrase was 0, innovative symbol was 1. So, I need to add 0 1 to the dictionary. So, I add 0 1 to my dictionary, and this is my codeword number 4; and 0 1 can be written as concatenation of 1 0 which is codeword 1 with 1 so that is why I am writing it in like this.

Next, so 1 1 2 has been decoded, next is 4. So, what is 4. If I look at my dictionary 4 corresponds to 0 1, so this 4 to be decoded as 0 1. So, this is decoded as 0 1. Now, what is the first bit of this sequence 0 1 that is 0. So, my innovative symbol is 0 and the previous source is 1. So, I am going to add 1 0 to my dictionary. So, then I add 1 0 to my dictionary that is code word number 5. And again this can be written as concatenation of 1 which is code word number 2 with this 0 is a symbol 0, so I can write it as concatenation of code word 2 with innovation symbol 0.

Next, I decode 5, what is 5, 5 is 1 0. So, 5 is decoded as 1 0. So, this is decoded as 1 0 that is this is 1 0. Now, in 1 0 what is the innovation symbol that is 1 this is the innovation symbol. Now, I add this innovation symbol to the previous phrase. So, I get 0 1 1. So, I am going to add this phrase 0 1 one to my dictionary and this will be my

codeword number 6. So, 0 1 1 can be written as 0 1 concatenated with 1 and what is 0 1 0 1 is my codeword number 4. So, code word number 4 concatenated with innovation symbol 1 so that is the dictionary entry number 6.

Now, the next received codeword corresponds to binary coding of 2. So, what is 2, 2 is just 1. So, this 2 will be decoded as 1 so that is 1 here. Now, the innovation symbol here is 1 the previous string was 1 0. So, I am going to add 1 0 1 to my dictionary. So, I add 1 0 1 to my dictionary and that is codeword number 7. And this can be written as coded number 5 which is 1 0 concatenated with 1 so that is my in the alternative presentation I am writing it as 5 concatenated with innovation symbol 1.

Next, I decode 5, what is 5, 5 is 1 0. So, this is decoded as 1 0, this is 1 0. Now, what is innovation symbol the innovation symbol is 1, previous phrase is 1, so add innovation symbol here. So, I am going to add 1 1 to my dictionary. So, I add 1 1 to dictionary that is codeword number 8, and this can be written as concatenation of second codeword with this innovation symbol which is 1.

Next, I will try to decode 3, what is 3, 3 corresponds to 0 0 so then this is decoded as 0 0 and this is my 0 0. Now, what is the innovation symbol here then innovation the layer is 0 and the previous phrase was 1 0. So, add this innovation symbol here and add this new phrase to my dictionary. So, I am going to add 1 0 0 into my dictionary and that is my codeword number 9 in the alternate form this can be written as concatenation of codeword number 5 which is 1 0 concatenated with innovation symbol 0.

Next 6, where 6 corresponds to 0 1 1, this will be decoded as 0 1 1, so that is what it is, it is 0 1 1. Now, innovation symbol here is the first that which is 0 previous phrase is 0 0. So, add 0 here. So, I am going to add 0 0 0 as my new phrase here that is my codeword number 10, which can be written as concatenation of codeword 3 with an innovation symbol 0.

Next 11, here is 11, I do not have 11 in my dictionary. What should I do now, I do not have my dictionary is only two 10 now. So, how do I decode 11, now this situation happens, if you go back to the encoding of the same sequence, this situation happens if we are going to use a string which has been put in the dictionary. If you look at this point 0 1 1 has been used as input in the dictionary; and immediately after that, we are using this 0 1 10. So, if you use a string which has just been put in the dictionary you encounter

situation like this where a string appears which is not there in the dictionary. So, 11 occurs, but 11 is not there in the dictionary, but this will happen if your string which has been just been added to a dictionary has appeared. So, this one should start with 0 1 0 1 1 and this bit will be nothing but the first bit of this. So, this string 11 would be 0 1 1 0.

Now why 0 1 1 0, as I said this situation happens when a string which has been added to a dictionary is been added to a dictionary is been immediately used. So, this entry number 11 codeword 11 should be at the form 0 1 one something innovative symbol. Now, what is this innovation symbol innovation symbol is nothing but the first bit of this new string and the first bit of this new string is 0. So, this bit and this bit here it should be same and that is why I am getting that 11 should be 0 1 1 0. So, I put 11 as 0 1 1 0 and I decode this as 0 1 1 0.

So, if you encounter a situation where a dictionary entry codeword has come up not there in dictionary at least you are using as phrase which has just been added to a dictionary. So, in this case, basically we had this 0 1 1 which is phrase I just occurred. So, the new entry should be of the form 0 1 1, I send new symbol. Now, this new symbol has to their first bit of the next decoded sequence. So, then innovative symbol here as to be 0 that is why this string corresponding to this code word 11, it should be of the form 0 1 1 0. So, we are decoded this.

Next, we go to 8 what is 8 that is 1 1. So, we decode 8 as 1 1. Now what is innovation symbol here innovation symbol here is 1. So, we add this to this old string which is 0 1 1 0 and then this is added to the dictionary. So, new entry in dictionary is 0 1 one 0 1 and that is my codeword number 12. This can be written as concatenation of codeword 11 with this innovation symbol 1. And finally, this 10 is nothing but 0 0 0. So, 10 is decoded. So, this is decoding of 8 and 10 is decoded as 0 0 0, and here the innovation symbol is 0. So, we add this here.

So, next entry here the dictionary would be 1 1 0 because the old phrase is 1 1 that is this and this innovation symbol here is 0. So, the entry layer would be 1 one 0 and this can be written as concatenation of 1 1 which is codeword number 8 with this innovation symbol 0. So, this can write like this. So, note that then this particular sequence has been decoded into string like this. So, with this, we conclude our discussion on Lempel-Ziv Welch algorithm.

Thank you.