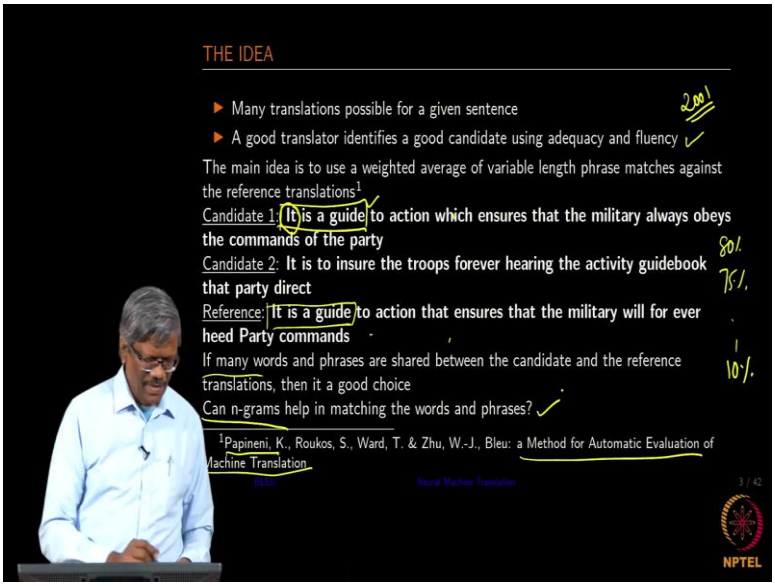
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture - 74
BLEU - "A short Discussion of the seminal paper"

(Refer Slide Time: 00:15)



Alright. So, as we know that there are many translations possible for a given foreign sentence right and then there is a set of references that are also available for the same sentence; for the same English sentence I should be able to construct a similar set of English sentences. So, I should be able to generate more English sentences which means the same thing correct.

And, then when you give that hypothesis to a human person, he immediately looks are it and then says this particular one is good; let us use that as the right translation for the foreign a sentence that we have seen ok. So, how does he do that? So, is he really looking at the sentence word by word or phrases or he looks at the entire sentence as a whole and then make sure that what he is seeing is the right translations? So, what is the mechanism by which he really translates and that is what we want to use to automate this?

So, one way is very simple this is a very simple assumption that we can do, look at every word from the hypothesis and then compare it with the reference sentence one by one.

And, if they match exactly the way in which the reference sentence is written then you pass that, otherwise, say that it has got some errors and do not pass. So, that is a very penalty way of doing it. So, we should be able to pass sentences there are at least 75 percent right or 80 percent right something like that.

Can we use some mechanism of what we had studied earlier right in terms of breaking the sentence into unigrams, bigrams and trigrams or 4 grams? And, then compare every word every n-gram with the reference sentence and then start providing a score as I mentioned earlier during the introduction of this right. So, here what I am going to be doing yes I am going to be providing two candidate sentences and then one reference sentence ok. One thing forgot to mention the idea that I am talking about is coming from this paper written by Papineni and others ok.

So, this is the one which is the seminal paper that you want to read as well if you want to really understand what went through when they were developing the standard. I am going to be doing some portions of the paper in this particular session as well ok. So, this came I guess in the year 2001 or 2 I am not sure. So, what they have done is there use the very simple idea of comparing the n-grams ok. And, then I send when they started doing that they also found that there are issues with respect to short sentences.

For example: if I just use these 4 words as my candidate sentence and then this is my reference here which also has those 4 words and this is the sentence that came out of the translation system. So, how do you validate this? So, I would like to say that these sentences 100 percent right ok. So, in a way you can score, but it is inadequate in the sent that the reference words are longer rather the reference sentences longer than the candidate one.

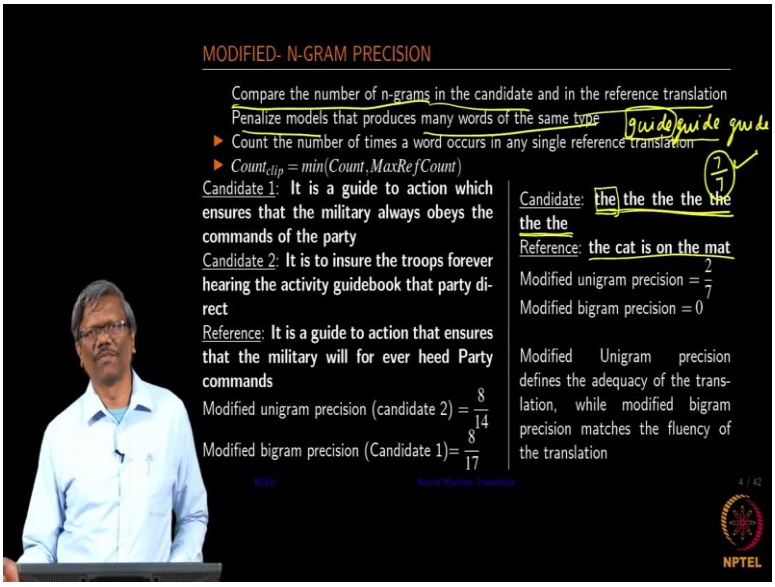
So, if I just provide you this particular word alone then it will pass ok, it is not the right sentence, it is just a single word right. So, we do not give a 100 percent rating to that ok. So, these rules taught coming out as in when there is started building this a standard and then we see how they really added those rules into the system to finally, create a good standard. So, as we look at the sentence right we start looking at that in terms of n-grams as we mentioned earlier.

And, then when we see more n-grams or found in the reference sentence as well, then we can say that this candidate has passed with a certain score; let us say about 80 percent or

75 percent and things like that or it is only about 10 percent ok. So, this is one way you know just you will found out find out how many unigrams are there and then how do you find this set of these right, from the set of these two. When you do the intersection you know the common words are coming out of these, right.

And, then if I count the number of words that are found due to the intersection and then if it is equal to the same number as the reference sentence, you have a 100 percent marks. If you do not have the numbers as in the reference sentence then the percentage starts coming down slowly, you from the 100 percent to maybe 5 percent, 10 percent; depending on how many words are aligned to the reference in terms ok. So, that is one very simple way of looking at that ok. So, that is why you know we going to be looking at the n-grams and see how those matching of n-grams are really helped us in terms of creating this standard ok.

(Refer Slide Time: 06:31)

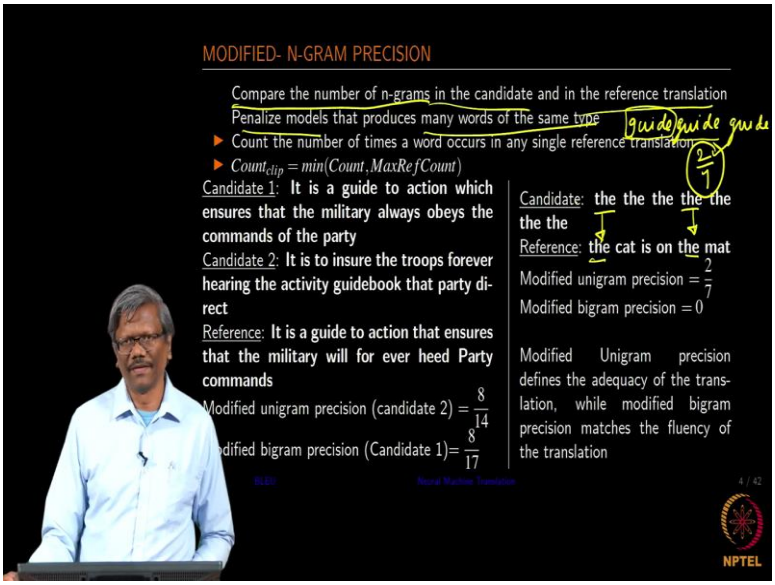


So, as I mentioned compare the number of n-grams in the candidate and in the reference translation it is that is one way and then penalizes models that are producing many words of the same type. For example, in this case, if my system creates something like this for example, I have a unigram model ok. And, then it looks at the language model and then sees I have seen guide, what is the next word and then based on the frequency it starts giving me the same thing. So, for example, this word is found very often.

So, it is possible that we can have a sentence of this type ok. We know that it is not the right sentence, but the candidate sentence could come from the translation system may have one of these ok. And, then we want to have a reference system as the cat is on the mat. So, the is found in this and it matches all the 7 words that are there in the reference system in terms of numbers ok. So, so we can say that when you match this the is found and then all 7 words that I have seen in terms of the count of reference. So, my unigram count could be 7 by 7, you know that it is wrong to correct.

So, there are only two places where this is connected to the reference one here and one here. So, when you won't use a good precision mechanism what you do is you just count how many times those words are available in the reference systems.

(Refer Slide Time: 08:37)



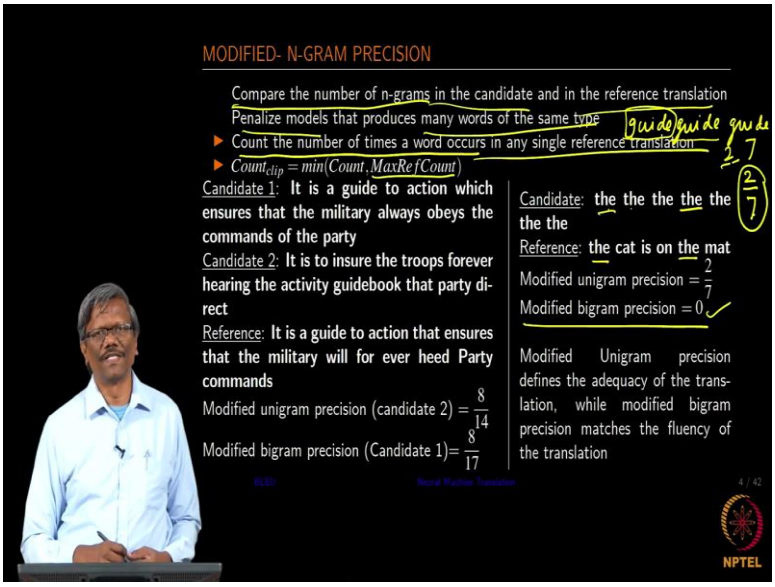
And then use that number and then find out how many words that your hypothesis contains and then use that as the denominator and the equator ratio of this time ok. So, this is the first way of looking at it. So, this is called the modified n-gram precision instead of scoring at 7 by 7 using the normal n-gram, we say that you are going to be looking at how many words are aligned or found in the reference system by looking at the candidate sentence. And, then use that for your precision computation ok.

So, in this case you look, I am just taking a very small short sentence for us and then we will go on to the long sentence later. So, I am going to be looking at the bigram. So, when you want to do the bigram; so, this is what the bigram is right. So, we want to be

looking at whether we have the and the happening in this, we do not find any of this. So, the modified bigram precision will give you 0, even though you have many bigrams available with respect to the count in the candidate sentence. And, then there is no bigram that matches with the reference bigram that you have correct.

So; that means, you have modified bigram precisions is 0, you go to the trigram again in the same fashion you can look at it when you have do not have the bigram there is no point in doing the other one, but still for the sake of understanding I am doing this ok. So, in this case, again there is no trigram available. So, your score is going to be 0, that means this candidate will get a score of 0 whereas, when you use the unigram you get the score of 2 by 7 ok. Now, look at the rule that we have, we count the number of times a word occurs in any single reference translation and, then use the reference count to clip that.

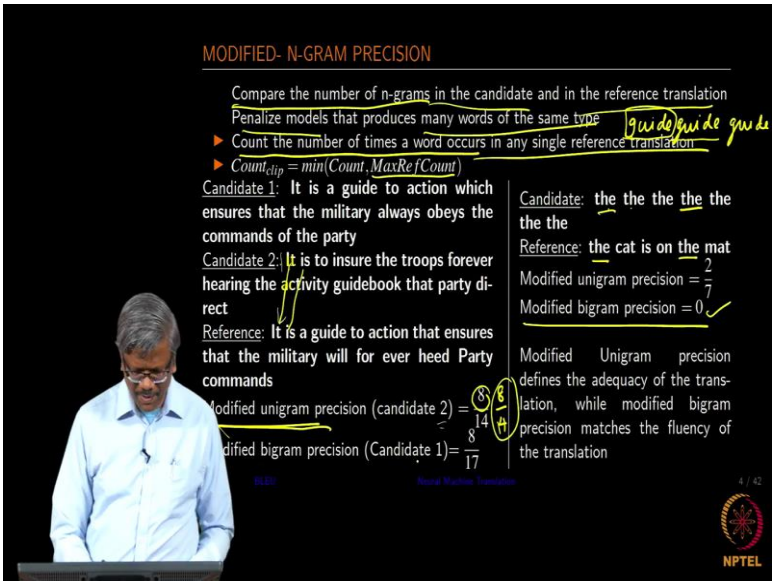
(Refer Slide Time: 11:19)



See for example you have 7 here and what you get is 2. So, now the sentence instead of 7, it becomes 2. So, you are going to be clipping that which says you go and then find out how many references you find in the reference sentence for the candidate words that you have here and then use that as your count, we call it as clipping the count ok. And, then use the total number of words in the candidate sentence, now we will let us go to the long one here.

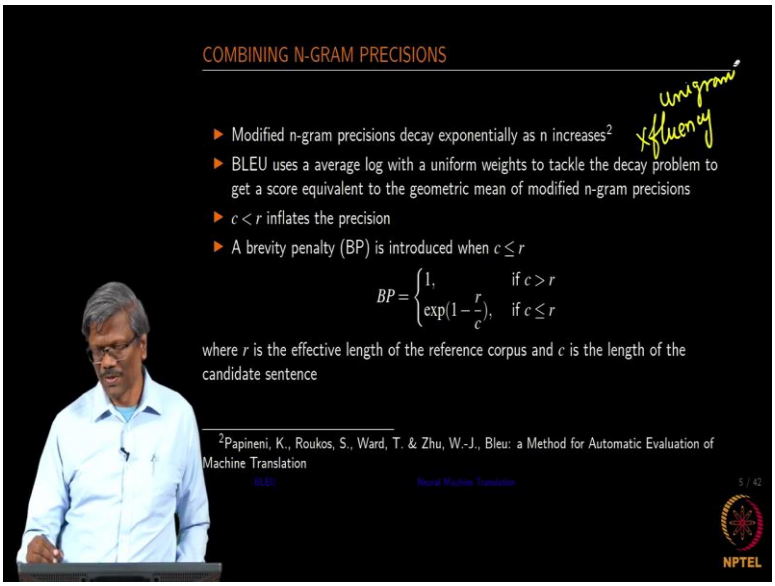
So, I am going to be looking at the modified unigram precision of the second candidate as we have done earlier. So, what you do is you just start looking at the words that are matching ok. So, in this case, you will find there are 8 words that are matched from the candidate 2 with the reference ok. And then how many words you have in this? There are about 14 words.

(Refer Slide Time: 12:49)



And, then use your modified precision as 8 by 14. So, if you look at the bigram procession for candidate 1, you will find 8 bigrams there are matching the reference sentence. And, then there will be 17 bigram counts in the candidate sentence and the procession becomes 8 by 17 ok. So, this is how you calculate the modified n-gram precision alright.

(Refer Slide Time: 13:27)



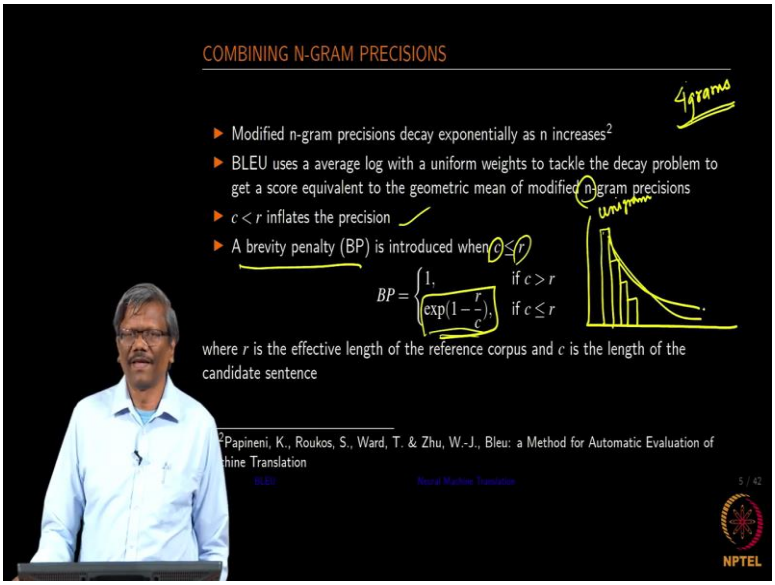
So, by combining all of this you remember we mentioned earlier that, it is not just enough to use unigram precision; you also have a look at the other n-gram precision. And, then combine them and provide a single score. So, why is that important to combine bigram and trigram or 4 grams? It is enough if you do with unigram. So, when you do that with unigram what is going on happen is you have actually doing it for adequacy right. So, we do not know whether the sentence is truly syntactically right or not, we have only just compared the number of words that are matching with the reference sentence right.

$$BP = \begin{cases} 1, & if\ c > r \\ \exp\ (1 - r/c), & if\ c < r \end{cases}$$

So, we will never be able to figure out the fluency part ok, with unigram you will never be able to find the fluency part because the context is not attached to that ok. Because it is only looking at one word at a time, you remember when you add more and more words in the n-gram even during the language modeling part also in the neural net learning of skip-gram base model and c bow by s model. When you provide more context, we would be able to the fluency of that language right. So, when we use unigram we say what it is not good enough, it probably gives you the adequate it will tell you whether the translation is adequate or not, but not beyond that.

So, that is the reason why we are now combining unigram, bigram, trigram, and 4 grams. So, in the case of BLEU 4 grams are used, all 4 grams are used in the BLEU model ok.

(Refer Slide Time: 15:25)



So, it is not just enough to compute all the modified precisions for all the 4 grams, we need to look out whether the sentences are short or long. You know if it is long I think there is already some penalty provided ok, it will be you will not find many pattern the matching or words that are matching. So, there is already an inbuilt penalty for the long candidate's sentences. For short sentences as I mentioned earlier there is no penalty, it would provide a good precision or then it would provide a good adequate translations score. So, we need to penalize the sentences that are pretty short as well.

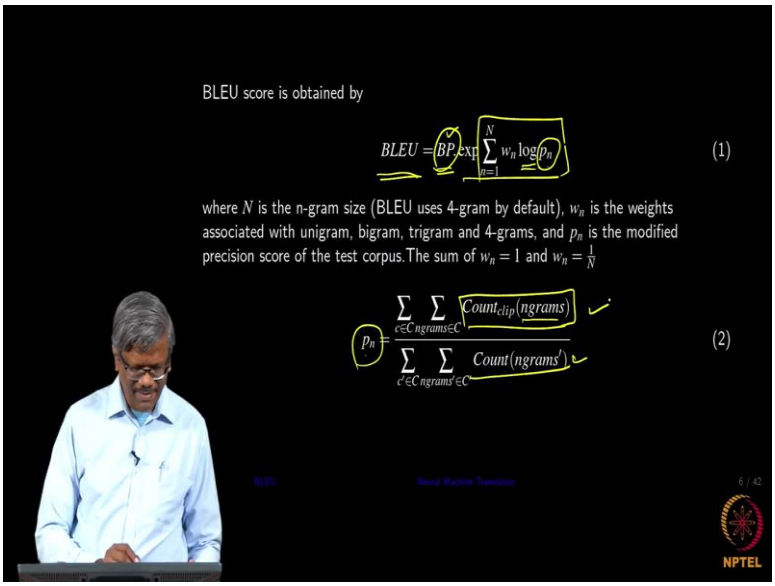
$$BP = \begin{cases} 1, & if\ c > r \\ \exp\ (1 - r/c), & if\ c < r \end{cases}$$

So, what BLEU does is it provides a brevity penalty call BP which is introduced; when the length of the candidate sentences less than or equal to their reference sentence we provider exponential decay for that ok. So, actually when you start looking at the translation score from sentence ok, using some small application; you will see that the unigram precision will be around here and then bigram would be here, trigram would be

here. So, if you add more and more sentences you will see that it is decaying in this fashion exponentially ok, as you add more number of n there ok.

So, to adequately compensate for small sentences which would give you a very high score right. So, this kind of exponential decays studied and incorporated as part of the brevity penalty for all the score that you are going to be computing here ok. And, it uses in expression BP equal to 1, if candidate sentences longer than the reference sentence, if it is smaller then we use the explanation given here ok. So, r is the effective length of the reference corpus and c is the length of the candidate sentence right. So, again as I mentioned earlier this we are talking about the paper written by Papineni and others ok.

(Refer Slide Time: 18:35)



So, now finally, you know you need to combine all the n-gram precision and the brevity penalty to get BLEU scores. So, when you compute the brevity penalty and then compute all the n-gram or the modified n-gram sum all of them and then finally, get these scores. So, when you do that we need to know what this as well ok. So, this is the modified precision score for the entire corpus. So, one thing we need to understand is BLEU works not with one sentence, it works with the combination of several reference statements ok.

So, usefully BLEU is not applied to one translation rather one sentence translation. So, you need to have a lot of reference translation available for you to really compute the score; so, that the reason we have this. So, in this case what we do for every candidate

sentence that we find we count the number of clips as we have done earlier and then sum them up using the n grams if it is 1 gram, unigram, bigram, and so on. And, then I use this in the denominator to find a score. So, this is done for the entire combination of candidate and reference translated, reference, or the gold standard that we have ok.

So, this clear. So, the way we are combining Bleus, get the brevity penalty and then for all the sentences candidate sentences you find the clip value one by one and sum them up. And, then you also know how many candidate sentences that you have and then count the number of words in all the sentences and finally, create one p n ok. So, this involves both the reference counts as well as the total number of counts in the candidate sentence.