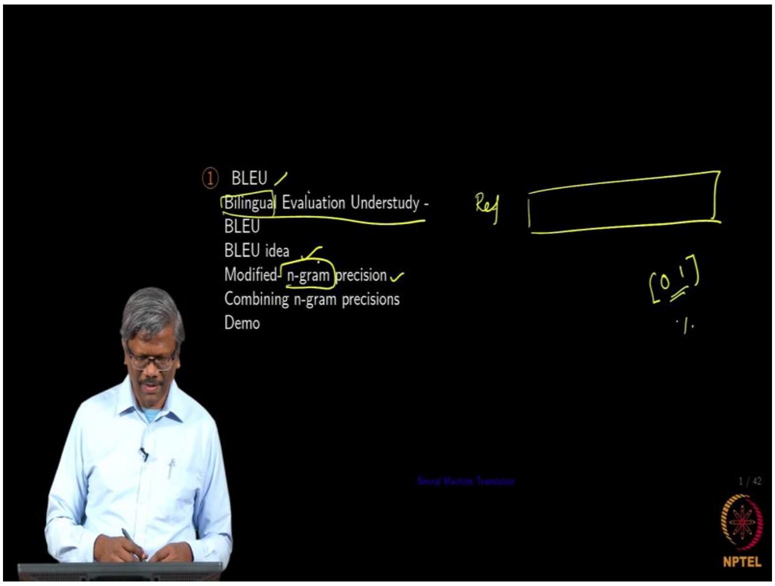
Applied Natural Language Processing
Prof. Ramaseshan Ramachandran
Department of Computer Science and Engineering
Chennai Mathematical Institute, Madras

Lecture – 73
Introduction to Evaluation of Machine Translation

(Refer Slide Time: 00:15)



1. BLEU ✓

Bilingual Evaluation Understudy - BLEU

BLEU idea ✓

Modified- n-gram precision ✓

Combining n-gram precisions

Demo

In continuation of our Machine Translation; one of the most important things that we have not done so far is the evaluation of the translated text right. So, in every application that you develop there is always a testing phase where you want to find out whether the application is working very well or not; is it not it? So, the developer usually what he does is he knows what is the input that requires for the application and starts with providing the inputs when you do the white box testing right.

You know what is the what is internal, so you start providing all the input to the application and then start testing them on various test cases. And then find out if every test case is successful or you give it to the black box testing, where you give the functionalities thereof the application and then those testers would test for application and then provide you the feedback which you will later come and then correct or fix the box and then move on ok.

This cycle continues until your application provides a very stable output for all the given inputs correct. In the same fashion when you develop an application in natural language processing, you also required these kinds of testing. If you had seen many apps right the kind of test cases that you prepare you to know may not be very huge. And the expected output is always known and they are not a very huge list of data points.

Whereas, in the case of natural language processing since it is possible for you to create sentences in all possible combinations right; meaning you would be able to innovatively create new sentences that semantically and syntactically give you the right meaning. We as we mentioned earlier natural language processing also provides lots of output for you. And then it is important for you to have a very reliable testing mechanism to make sure that your application is doing the right things so that you can move the application to the production.

To give an example in the case of the sentiment analysis you probably would have given so many different sentences to the application to make sure that it provides the proper sentiments of the input sentence that it has. It also should be able to provide the right sentiment for the sentences that it has not seen earlier, but you would have seen all the combinations of those words and things like that.

So, based on that it gives you some reliable output, and the tester also make sure that based on the certain gold standard that is set for those new sentences; you would compare it and then make sure every case has been passed and finally, pass the application to the production site.

In the case of machine translation it is again an important activity right. So, when the application generates a lot of candidates you need to test those candidates or the hypothesis that comes out the translation system. So, how do you do that? Every time when you make a small change, you want to make sure that you have not broken anything and the system outputs the right translation.

Is it possible for anyone to test all the hypothesis that comes out of the machine translation system; it is very difficult right even to hand-code all the manual translation we mention that it is going to be very hard. So, how are you going to really evaluate that translation that is coming out of the emission translation system? So, for about several years of people have been doing this in somewhat manual fashion; you know there is a certain mechanism they had, but until the late 80s or early 90s; there was any system that was available to really do the testing in.

Even though machine translation is around for 50 years; a good standard has not evolved. So, when IBM developed all those model they also found that there is a need to develop a certain automatic way of finding out whether translations are right or wrong. They started developing a few standards and one of the standards that came out of their stable is BLEU that is a Bilingual Evaluation Understudy is that the expansion for BLEU was ok.

In this what they try to do was how do I really compare certain hypothesis that is coming out of the translation system with some old standard that is already established. For example, you have pairs of sentences and then you know for that particular sentence; this is the expected translation. But when you use the translation system; there are several hypotheses that are coming out which is not going to be exactly the reference sentence that you have established.

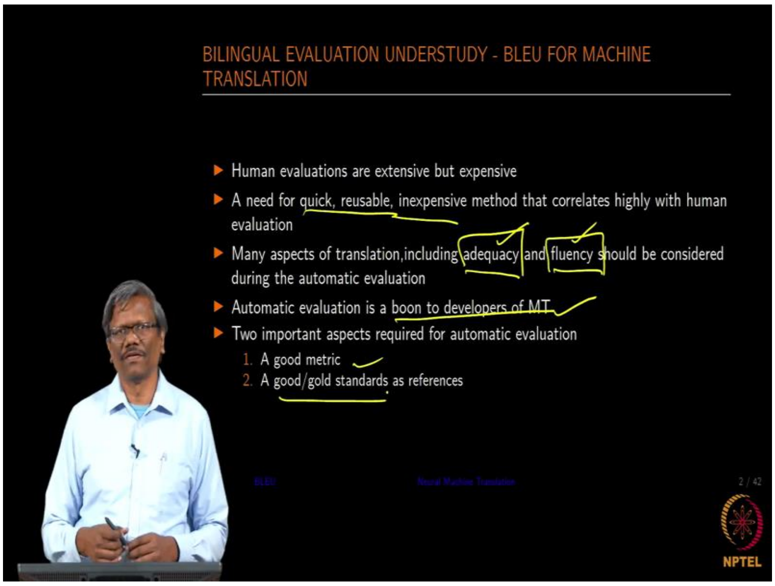
But you need to compare all the hypothesis that is coming out of those or all the candidates that are coming out of the machine translation system and then rate each one of them to make sure that you pick the right candidate translation right. So, for that they have started looking at the sentences and try to figure out what could be the possible ways that I could automatically compare two sentences ok.

So, in this session we going to be looking at the idea of BLEU and then we will look at how they modify the n-gram precision a bit. And then how do I compare or add or combine multiple n-gram precisions and then give one single score. I am sure you know about the n-gram right. So, when you have a sentence like this right.

So, the bigram evaluation and then assume that there is an equivalent reference that is available here. So, what bigram or the n-gram precisions do is we first start looking at 1 gram right. So, that is one word and then find out if it is there in the reference; take the second one it is there in the reference third one in the reference and so on. So, you do not really look at the order, but find out whether it is available or not ok.

And then create some score and say that this particular translation is giving you something like you know you can have a score of 0 to 1 or you sometimes you give the score in the percentage. If higher the percentage or higher the value in this range, you say that the translation is ok right. So, let us see how this goes. So, what I am going to be doing in these classes talk about the idea and then talk about the modified n-gram precision. And how do we combine the n-gram precision and then lateral I will just give you a demo for about a few minutes on how to really compare two sentences and then get a score all right.

(Refer Slide Time: 08:41)



BILINGUAL EVALUATION UNDERSTUDY - BLEU FOR MACHINE TRANSLATION

- Human evaluations are extensive but expensive
- A need for quick, reusable, inexpensive method that correlates highly with human evaluation
- Many aspects of translation,including adequacy and fluency should be considered during the automatic evaluation
- Automatic evaluation is a boon to developers of MT
- Two important aspects required for automatic evaluation
  1. A good metric
  2. A good/gold standards as references

BLEU Neural Machine Translation 2 / 42

Comparing; comparing the candidate and the reference is very expensive if you want to do it in a human way right. So, humanely it is an impossible task I will not say impossible task, it is going to be an expensive task right. So, you can just bring in 1000s of people and then ask them to verify the translation if they are good in both languages right. So, manually you can provide a 10 translation to each one of them and then if you have about million; you can divide that among those 1000 people and then ask them to verify the translation and then rate them right. So, that is one way of doing it.

So, it is going to be an expensive task and then it is not possible for you to do a real-time translation like we normally do with the Bingo or with Google right. So, in those cases it is not possible to give it to a human and say hey I am going to be talking to someone and this is going to be the sentence; I am going to be speaking tell me what is the translation that I will not be having in that language and then bring in another expect to understand that language and then ask him to verify all that is it is impossible to do that is not it.

So, what we want to do is to see if we can automate the task whether it is a batch-wise or in a real-time job. So, we need a very quick and reusable and inexpensive method that correlates highly with human evolution. So, when you do the translation; you want to make sure that it is pretty close to what humans will accept right as a good translation.

So, that is something that we want to provide as part of the automatic system. We also look at the adequacy and the fluency for example, if you go out to a place where you; you do not know the language of that particular local area and you know only English and the person that you are talking to other is also someone who knows his language well his native language well, but he understands few English words.

So, in those as situations what do you do? You just ask him whether he understand your native language and when he responds no and he says I know a few words in English. So, you start communicating with a broken English by giving the nouns and subjects right; not in the very semantic fashion. And then the person might understand that you know based on the understanding of the words that he has learned. So, you would be able to communicate using broken English just words without fray of forming a good syntactically right sentence.

And if he is able to respond back and then do what you want we can say that what you have communicated is good enough to talk to that person it is enough. So, those words are enough for from his perspective; he has some reference and his mind and then you have spoken certain words, he compares that and then says ok this is what this person is asking and he performed his job. The fluency is something that is what you are really speaking right; they were in a very fluent way with all those words connected syntactically and you are doing it also right in the semantic function.

So, we want to be able to achieve these two goals as well as part of the automatic evaluation. And then it is definitely a boon for the developers because he should not be divide depending on the human evaluators who would be writing this sentence and then you going to might be changing. As a developer you know you keep modifying it very fast and then you want to really see the result of your modification at the end of the; training phase ok.

And then you do not want to wait for a day or 2; you want a system that immediately gives you reasonable output using which you can start debugging or add additional (Refer Time: 13:26) to it so that your translations becomes better and better. So, this is a big boon for the developers initially and then we also want to have a good metric that is very consistent; see I do not want to be getting up different resultants for the same translated sentence at different points in time.

So, it should be a reliable metric for me. We also need a good gold standard as references; that means, we need a huge collection of reference translation that is available so that we can compare them with the hypothesis that we generate using machine translations systems right. So, this is the crux of why we definitely required this and then the result of all this is a standard that came out of IBM and they call it a BLEU. So, it is a bilingual evaluation understudy ok.