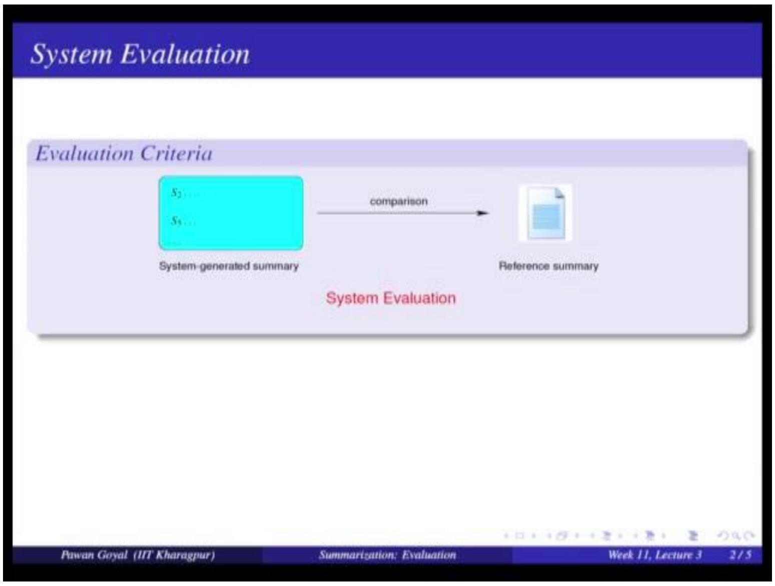
Natural Language Processing
Prof. Pawan Goyal
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture - 53
Summarization – Evaluation

Welcome back to the third lecture of this week, so we have talked about different approaches of summarization; in this lecture we are talking about evaluation. So, how do we evaluate the summary that you have obtained using your system, so what do I mean by evaluation.

(Refer Slide Time: 00:33)



System Evaluation

Evaluation Criteria

$S_2 \ldots$

$S_5 \ldots$

$\ldots$

comparison

System-generated summary

Reference summary

System Evaluation

Pawan Goyal (IIT Kharagpur) Summarization: Evaluation Week 11, Lecture 3 2 / 5

So, you run your summarization algorithm on your document set and then you are getting some set of sentences. So, as such assuming that you are running as extractive summarization; you will get a set of sentences as an output. You can also run different sort of abstractive summarization or whatever and you will get different sort of sentences as your output.

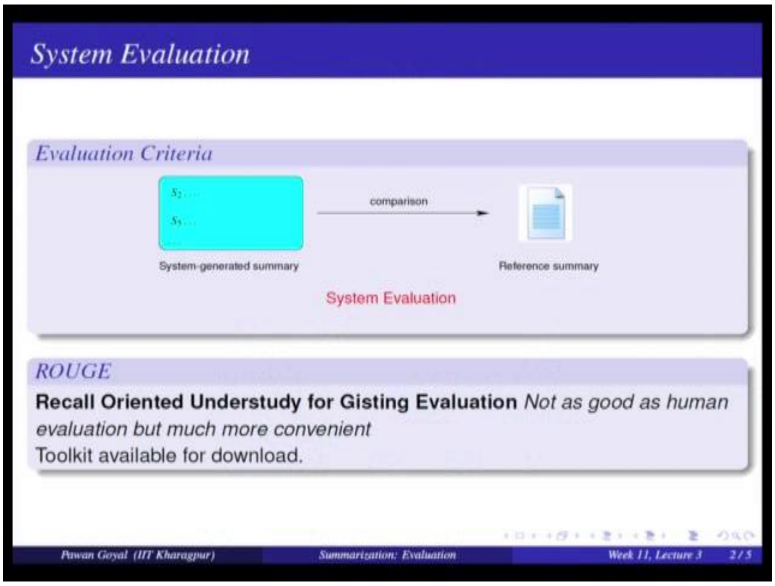
Now, how do I find out how good my summary is? So as such if you talk about, we will see what should be in criteria of evaluating a summary. So, we say if we give it to human, so human should say this summary contains the good amount of information from the original article. So, that is one criteria can be what is the informational coverage of this summary and then we can talk about how diverse the sentences are, so you are getting different sort of summary in here, you may talk about readability; if you are doing an abstractive summarization and there can be some other criteria depending on your what is your task. So, are you getting the information that you wanted from this summary.

So, this you can do by some sort of human evaluation, you give the summary to some humans and ask them to rate the summary along these points from a scale say 1 to 5 and it is how good the summary is on the scale from 1 to 5 as per the readability, as per the information coverage as per the diversity and so on. But what we will see is there some method by which you can automatically evaluate; how good the summary is with respect to the document and there are some methods for doing that and one very popular method is using the Rouge score and that is what we will see; what is the idea?

So, your system is rating the summary for this document now in Rouge's method; you will assume that there are some humans that have created a gold standard summary for this document already. So, this helps in that you can now try out different different systems; get the summary and find out which one does the best with respect to the human. So, you have the human summary and you have some system generated summary and you want to compare these and this comparison is done by an approach; it is called the Rouge approach.

So, what do we have? So we have the system generated summary here, some sentences and a reference summary that is created by the human. Now in general; it may happen that you want to create multiple summaries for the same document, so the same document might have multiple reference summaries; so 3 reference summaries or 5 reference summaries still you should be able to compare this with the system generated summary.

(Refer Slide Time: 03:26)



System Evaluation

Evaluation Criteria

$S_2 \ldots$
$S_5 \ldots$
comparison
System-generated summary
Reference summary

System Evaluation

ROUGE

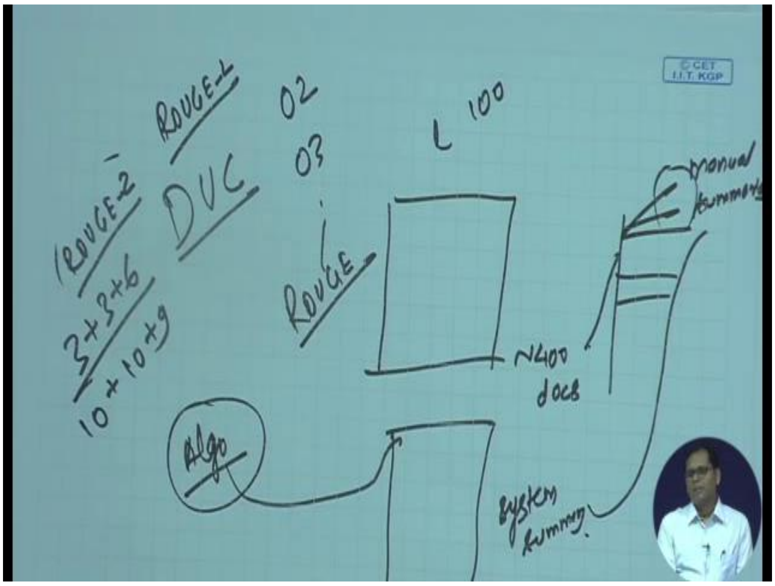
Recall Oriented Understudy for Gisting Evaluation *Not as good as human evaluation but much more convenient*

Toolkit available for download.

Pawan Goyal (IIT Kharagpur) Summarization: Evaluation Week 11, Lecture 3 2 / 5

So, the approach that is used is called the Rouge evaluation; recall oriented understudy for Gisting evaluation. So, as good as doing a human evaluation, but still it is quite convenient because for summarization, you have a lot of different computations and benchmark data sets. So, what you want to see that different systems as you propose different algorithms; are they able to improve on the previous approaches or not and a human evaluation may not be first of all feasible for doing for different algorithms, different variations and if it is there it may not be very reliable that do you get the same humans again and again and even if they are from different volunteers; how much they agree and all, so there can be many other issues with that. So, using an automated evaluation avoids all such problems and you say that for a benchmark data set. So, for summarization there are datasets like that are given in document understanding conferences.

(Refer Slide Time: 04:32)

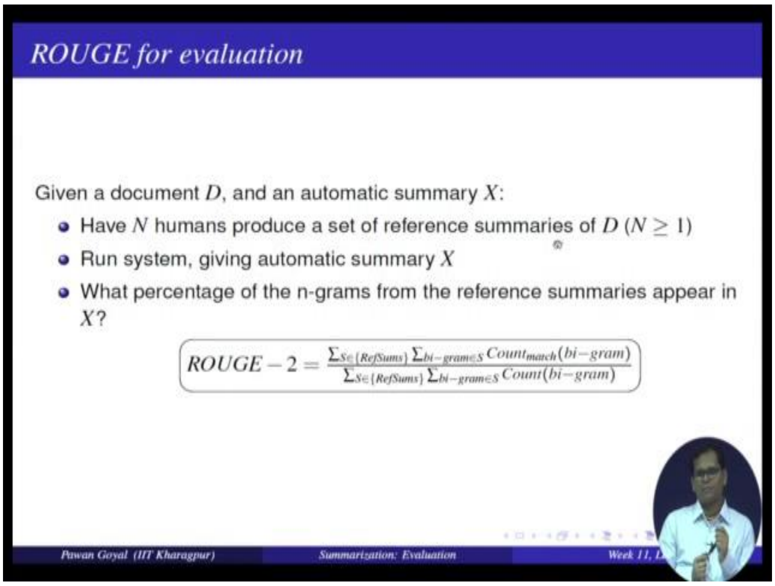


So, they are by the name of DUC, so DUC has multiple datasets, so DUC 02, DUC 03 and so on. So there what do you have? You have some documents, so like they may be roughly 400 documents; for each document they provide some manual summary. So, this is again by using certain guidelines; they have built the manual summaries. So, they have done it once for all these 400 documents.

So, now once you have your algorithm; you again produce the summary for these 400 documents. So, this is your human summary, this is your system summary and then you try to see how close is the system summary to the manual summaries and this one; this thing you will do for. So, now, you can vary your algorithm and see I will find the how good this algorithm works on this dataset, how good the other variation works on this dataset and you can also compare on this benchmark, so that is why this automated evaluation is helpful.

Now, Rouge is one very popular method for doing this, so let us see how what is the basic idea for Rouge. So, Rouge again has many different variations, so that is do you want to find out similarity on only the unigrams; that is single words how much is single words are matching or bigrams; how many bigrams are matching or you want to go for longest common subsequence that you are matching and like that there are many variations. So, most popular ones are Rouge 1 and Rouge 2; how many unigrams are matching and how many bigrams are matching between the 2.

(Refer Slide Time: 06:26)



ROUGE for evaluation

Given a document D, and an automatic summary X:

- Have N humans produce a set of reference summaries of D ($N \geq 1$)
- Run system, giving automatic summary X
- What percentage of the n-grams from the reference summaries appear in X?

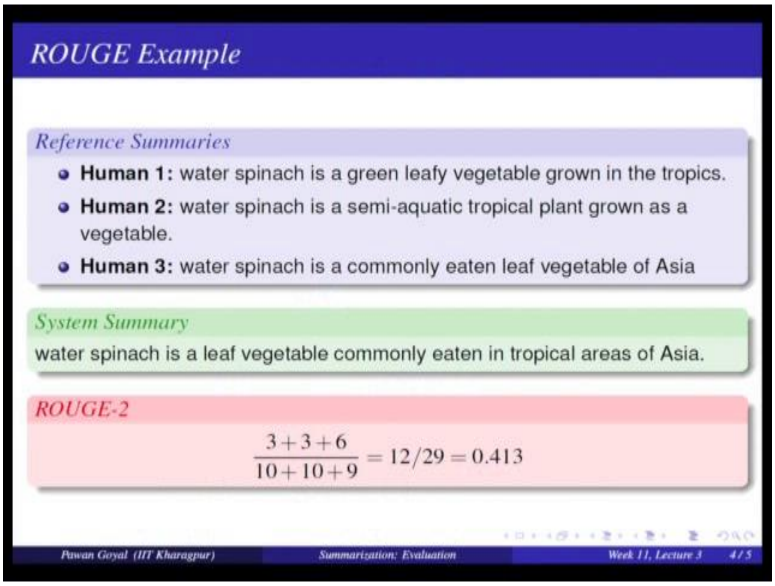
$$\boxed{ROUGE-2 = \frac{\sum_{S\in\{RefSums\}}\sum_{bi-gram\in S} Count_{match}(bi-gram)}{\sum_{S\in\{RefSums\}}\sum_{bi-gram\in S} Count(bi-gram)}}$$

So, what is the actual approach for doing that; so, suppose I have a document D and an automatic summary X; X is what my algorithm is providing. Now what I will do? I will have N humans produce a set of reference summaries. So, like I did in this (Refer Time: 06:43) dataset, so there are suppose 3 humans; they produce summary for each of this document. Now I have the automatic summary X, so what I will do? I will find out what percentage of the N gram from the reference summaries appear in X. So, this is like I am finding the record, how many N grams from reference summaries are appearing in the system summary.

So this N I can vary from 1 to n so on, so if I take n is equal to 1; this is the Rouge 1 measure; if I take n is equal to 2; this is the Rouge 2 measure. So, this is the example for using a Rouge 2 measure; that is you are taking bigrams. So, what you are doing? So you are counting; so your numerator is for all reference summaries; for all the bigrams that appear in the reference summaries; find out how many are matching with the system generated X; how many bigrams are there in the system generated summary x as well. So, you are counting that this is numerator; what is denominator, denominator is for all the reference summaries for all the bigrams count how many bigrams are there; that is you are finding out among all the bigrams that you can find in the all the summaries; how many are present in the human sorry system summary and immediately you can see that; if a bigram occurs in many of the human generated summaries, it gets a higher weight. So, if that bigram occurs in the system generated summary, it will start giving it a high weightage.

So, this is a very simple approach but this has shown to correlate a lot with the human judgments and this is one of the very popular method for evaluation, so this is well accepted method. So, let us try to see how this works on a simple example.

(Refer Slide Time: 08:33)



ROUGE Example

Reference Summaries

- **Human 1:** water spinach is a green leafy vegetable grown in the tropics.
- **Human 2:** water spinach is a semi-aquatic tropical plant grown as a vegetable.
- **Human 3:** water spinach is a commonly eaten leaf vegetable of Asia

System Summary

water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

ROUGE-2

$$\frac{3+3+6}{10+10+9} = 12/29 = 0.413$$

Pawan Goyal (IIT Kharagpur) Summarization: Evaluation Week 11, Lecture 3 4 / 5

So, suppose I have a document and there are 3 reference summaries provided by 3 humans. So, first summaries water spinach is a green leafy vegetable grown in the tropics, second one water spinach is a semi-aquatic tropical plant grown as a vegetable and third water spinach is a commonly eaten leaf vegetable of Asia; 3 different reference summaries and now your system produces this summary, water spinach is a leafy vegetable commonly eaten in tropical areas of Asia. So, what you will do? You will find out how many bigrams are common in human summary one and system summary plus common bigrams in 2 and summary plus common bigrams in 3 and summary that will be your numerator, denominator will be bigrams in 1 plus bigrams in 2 plus, bigrams in 3.

So, what are the bigrams in 3, so bigrams will be 1 less than the number of words, so because I can take bigrams like water spinach; spinach is a and so on. So, number of words are - 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, so there are 9 bigrams in third summary. Similarly here 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11; 10 bigrams, so my denominator here is 29; now what will be my numerator? So, numerator would be how many bigrams are common from here to each of these, so let us see system summary and human one water spinach is common 1 spinach is 2 is a 3 then system and human 2 water spinach 1; spinach is 2 is a 3 and then nothing else no other bigram occur in both. So, 3 plus 3 and system and human 3 water spinach 1 spinach is 2 is a 3, then commonly eaten 4 then leaf vegetable 5 of Asia 6.

So, number of common bigrams are so 3 plus 3 plus 6 divided by total bigrams 10 plus 10 plus 9 that is the Rouge 2 score. Now this you will compute this you have computed for what this you computed for this one document, we will compute it for all the documents in your corpus. So, there is a toolkit available for Rouge, you can download that toolkit set up your system and then you will give all the documents; it will compute all the scores Rouge 1, Rouge 2, there is Rouge l and so on. There are many many other scores that you can compute and it can also tell you the confidence interval and all and then you can compare different approaches; how much they are the Rouge scores are different.

Now, one thing you must be careful while using the Rouge approach; in general when we are doing it for benchmark datasets, we have some constraint that I want a summary of length 100. But suppose when you are running your system some sentences are having, so you are stopping when the length is just above 100. So, suppose you are stopping at 100 and 500 and 700 and 8. So, if you want to use those summaries also there is a way in which in the toolkit you can provide this say that I have the length I should consider only length up to say 100 of a given system summary.

So, there is a parameter you can set in the toolkit that say I will take only up to 100 words will be taken everything else will be left out, so that is possible. So, by the same Rouge summary, you can choose different different lengths, you can also say that I want to do stemming. So, stemming will help in that words like boy and boys will be starting matching; if boy occurs in reference summary and boys occurs in the system summary; it will not match here but if you say I will do stemming then it will start matching. You can also say that I will remove the stop words before completing all this. So, that way you are only matching the non-stop words, not like is and a occurring in both that will not make sense.

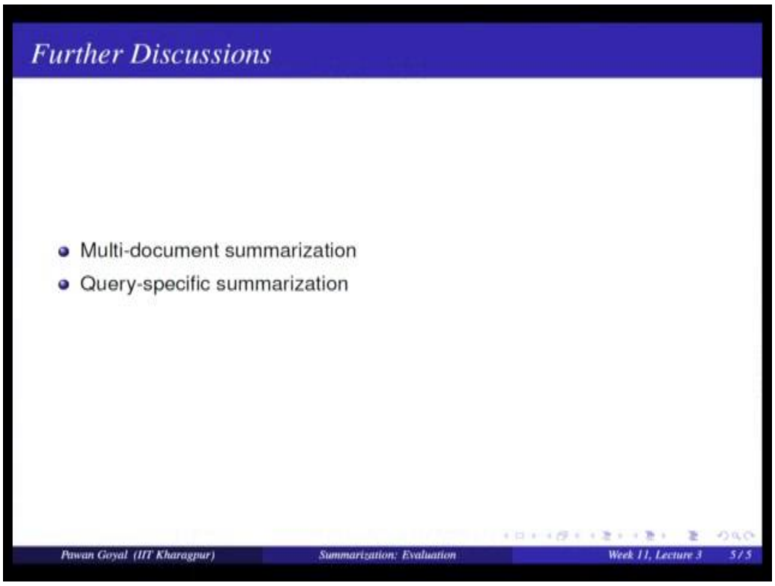
So, there are many other options that you can see in the Rouge toolkit and they help you in getting a very good evaluation of your system. So that is what we saw here Rouge 2 score, if you compute for this these 3 summaries this particular system summary it will come out to be 0.413 and you can compute Rouge 1 and so on like this by this method.

So, there are many other methods for evaluation, so one other method is like pyramid evaluation. So, where the idea is that you define some sort of semantic content units from your documents but they have to be manually defined and that is why it does not become completely automatic, if you have a large corpus you have to define by on your own some semantic content unit and then your idea will be whatever summary you are getting, you should have the maximum coverage of those semantic units in a nice, so this is a nice method also but it is not it requires you to define these units and there are other approaches that have seen that. So, given the document and the summary can you find some property among these that can tell you whether this is a good summary or not but still I would say you can always rely on Rouge 2. If you have a ground truth summary if you do not have ground truth summary one way is you create your own summary or you do you go for some human evaluation if you do not have a lot of documents to evaluate.

So, that finishes our discussion on summarization, but so there are certain things that we did not discussed. So, this is something that we talked in the initial one of the initial slides that what are the different genres of summarization. So, we have focused our whole discussion on single document summarization and generic summarization and extractive summarization, what we were doing taking the sentences extracting the sentences but what about doing a multi document summarization.

So, multiple document summarization is not very different from what we are seeing here, except that now you are having different documents. So, as such you can apply the same algorithm like lexing algorithm where, you take all the sentences in all the documents and apply your algorithm but there once you have the summary you might want to give different weightage to different documents. Suppose you know one document is more important than another document, you will have more weightage for that sentences from that document. Also the similarity between the sentences in same document might be given different sort of function and similarity across documents, you might also use how many how similar two documents are and so on. So, there are many different tricks that you can apply there but overall the idea is roughly the same, so it would not be very very different.

(Refer Slide Time: 15:43)



Further Discussions

- Multi-document summarization
- Query-specific summarization

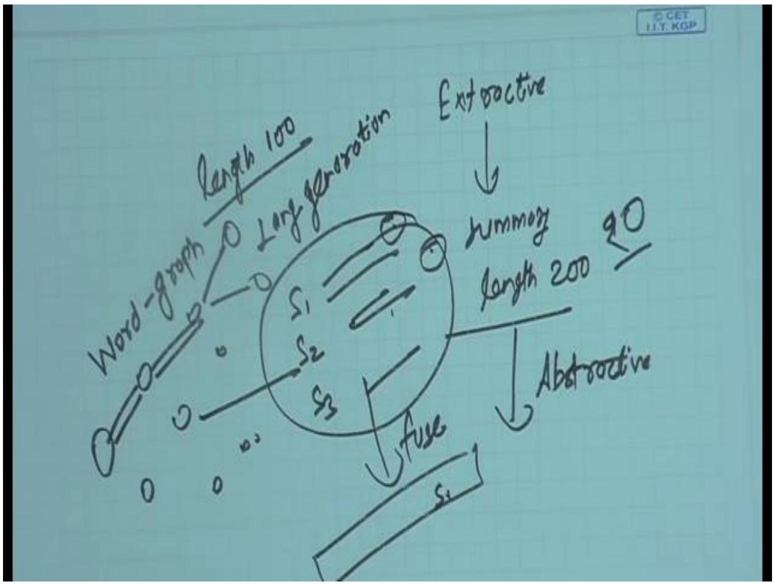
Pawan Goyal (IIT Kharagpur) Summarization: Evaluation Week 11, Lecture 3 5/5

Then there is something like a query specific summarization; that is user has given a query and I want a summary with respect to that query. So, one particular approach would be first you find out what are the sentences that are similar to that query and then you summarize only on those that way reduce your initial set or it can be first you get a larger summary than what you wanted and then you apply your; you find out which ones are more closer to the query then you retain only that.

So, first find out the important sentences from there retain only those that are closer to the query and there can be some other methods also. So, idea is and here again you can do a query specific multi document summarization. So, you have a query, you have a whole repository; first find out using some information to the method which of the documents are relevant to the query, so this becomes your multi document set from there using your summarization and pick those sentences that are relevant to your query. So, this is like you can apply query especially summarization by using same sort of techniques but some preprocessing some post-processing based on the query.

And then finally, you have the abstractive summarization, now abstractive is slightly trickier. So, because here now you are not just worried about just picking up the sentences, you are taking the sentences if some sentence contains similar information you are trying to fuse them together. So, what is in general done is that whenever you want to produce abstractive summary, you apply the extractive summary first get a larger set.

(Refer Slide Time: 17:19)



So, that is suppose I want a summary of length 100 I want a summary of length 100, so what I will do? I will apply extractive summarization get a summary of say length 200. So, twice more than what I need; now I apply some abstractive method in abstractive method in general what I will do. So, here I will take; suppose I have here 10 sentences 10 or say 20 sentences, I will find out which sentences are very similar or close to each other. Suppose I find S1, S 2 and S 3 are similar to each other.

So, what I will do? I will take these sentences and try to fuse them together to get a single sentence; then how this fusion will take place? I will say are there some words that are common here and there and then some information that is providing some information that this is providing. So, I will take the common part take information from here and here. So, there are many approaches for doing that many of them are based on a word graph.

So, that is you construct all the words, take all the words in the sentences, construct a graph and then you try to follow the path, these words are connected, these words are connected in the second sentence also these words will be connected, but there will be some diversion wherever there is diversion you try to take both of these together, by doing some sort of generation language generation. So, that is like you can connect them by and, or some other different connectors but whatever is common you will take it only once.

So, there are again, this is again a nice area where lot of work is happening that how do you construct the abstractive summarization. But again the idea is that you first take the extractive one and then create an abstractive from there and again here you can use your IIP based method to find out what paths to be taken, paths can be based on the unigrams bigrams or higher length. So, with that I will say that this finishes our discussion on summarization and this is a very nice application lot of work has been done and is being done right now in this field, there are lot of different applications in.

So, you can talk about news summarization and scientific article summarization at the same time you can talk about summarization from your feeder streams, quora answers, stack flow, stack overflow answers and so on. So, lots of tweets are there for certain events you want to summarize those from a disaster event you want to summarize those. So, that way you can use your extractive summarization abstractive summarization to also be able to help you.

So, in the next lecture we will start a new application topic that is text classification again that is very very important and we will discuss one very again very appropriate baseline for that.

Thank you.