AN AUTOMATIC TEXT SUMMARIZATION USING LEXICAL COHESION AND CORRELATION OF SENTENCES

A.R.Kulkarni[1], S.S.Apte[2]

[1] *Computer Science & Engineering Department, Walchand Institute of Technology, Solapur – 413006, India*
[2] *Head, Computer Science & Engineering Department, Walchand Institute of Technology, Solapur – 413006, India*

Abstract

Due to substantial increase in the amount of information on the Internet, it has become extremely difficult to search for relevant documents needed by the users. To solve this problem, Text summarization is used which produces the summary of documents such that the summary contains important content of the document. This paper proposes a better approach for text summarization using lexical chaining and correlation of sentences. Lexical chains are created using Wordnet . The score of each Lexical chain is calculated based on keyword strength, Tf-idf & other features. The concept of using lexical chains helps to analyze the document semantically and the concept of correlation of sentences helps to consider the relation of sentence with preceding or succeeding sentence. This improves the quality of summary generated.

In this paper we discuss a summarization method, which combines lexical chaining with correlation of sentences in which relation of a sentence with the preceding sentence is considered. Our experiments show that the inclusion of both these features improves the quality of summary generated.

Keywords— *Text summarization, Wordnet, Correlation of sentences, Lexical chains*

---***---

1. INTRODUCTION

1.1 Motivation

These days, the number of Web pages on the Internet almost doubles every year as the information is now available from a variety of sources. It takes considerable amount of time to find the relevant information. Automatic Text Summarization will help the users to find the relevant information rapidly. It generates the summary of the document and one can read the read the summary and decide the relevance of the document to the information needed by the user.

1.2 Background Research:

Text summarization is the process of producing a condensed version of original document. This condensed version should have important content of the original document. Research is being done since many years to generate coherent and indicative summaries using different techniques. According to (Jones, 1993) the text summarization is described as two step process

i) Building a source representation from the original document.
ii) Generating summary from the source representation

Text summarization can be broadly classified into two types: Single document summarization and multi-document summarization. This paper focuses on single document summarization that generates summary of single document. The text summarization can be categorized into extractive and abstractive based on the nature of text representation in the summary.

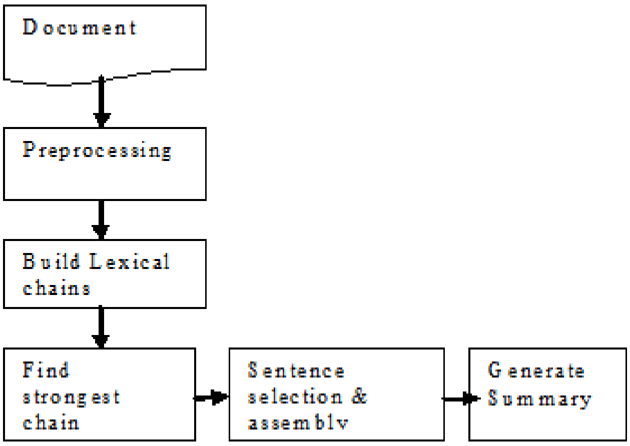
Many methods have been proposed till now on generating a coherent summary. The earlier methods used only statistical methods that focused on term frequency [1] for choosing important sentences. These methods were not found to be efficient as it did not consider all the contexts of the word or identify semantically related terms known as cohesion.

Then came methods which used semantic representation of the original document supported by a domain-specific knowledge base. Now a days text summarization is considered as a natural language processing task . Lexical chains a simplest form of lexical cohesion was introduced by Morns & Hirst[2].But it was found that all possible senses of the word were not taken into account. .

Berzilay & Elhada [2] presented a better algorithm that constructs all possible interpretations of the source text using lexical chains. It is an efficient method for text summarization as lexical chains identify and capture important concepts of the document without going into deep semantic analyses. Lexical chains are constructed using some knowledge base that contains nouns and its various associations.

Our Algorithm is based on the method used above. We have used Wordnet to generate domain-specific extractive summary using Lexical chains for the nouns in the document. The algorithm segments the given content into sentences & then into tokens. These tokens are tagged using POS tagger. The Nouns are selected & for each noun in the segment, we consider its sense using Wordnet. Then we attempt to merge these senses into all of the existing chains in all possible ways, hence building every possible interpretation of the segment. Next merge chains between segments that contain a word in the same sense in common. The algorithm then calculates score of lexical chains, determines the strongest chain and uses this to generate a summary. We have also used the concept of correlation of sentences to generate a good quality summary.. The terms that occur in the strongest lexical chains are considered as key terms and the score of sentences is calculated based on the presence of key terms in it. All the sentences are ranked based on their score and top n sentences are selected for inclusion in the summary. Then the correlation of sentences is checked and if any sentence has correlation with the previous sentence, then the previous sentence should also be included in the summary based on condition as shown in the algorithm below

2. ARCHITECTURE OF TEXT SUMMARIZATION



Preprocessing includes

- Segmentation
- Tokenization
- POS(part of speech tagging) at lexical level.
- Stemming.

3. LEXICAL CHAIN COMPUTING ALGORITHM

1. Input Original document for generating summary (.txt file).
2. Divide the document into sentences using segmentation.
3. Each sentence is divided into tokens using tokenizer.
4. These tokens are tagged using POS Tagger.
5. For each noun build the synsets.
6. For each sentence generate a map using 4 relations: Synonym, Hyperrnym, Hyponym, Merynym.
7. Calculate distance of each word from other related words.
8. Build Lexical chains using generated map.
9. Calculate each chain weight using values of distances of each word
10. Select longest chain i.e. best chain having highest chain weight
11. From the original document select sentences that have words in the best chain retaining their order of occurrence in the original document.
12. Pick top n sentences as summary based on the percentage of original document to be used for generating summary.
13. If the selected sentence starts with words : although, however, moreover ,also, this, those and that ,then they are related with the preceding sentence.
14. If the rank of the preceding sentence is equal to or greater than 70% of the rank of the selected sentence, then it is included in the summary. In this way correlation between sentences is maintained.

4. EVALUATION

Evaluation is the most important part of any research work. It helps to compare various techniques based on evaluation metrics.

This paper uses precision & recall [4,5,6]technique for evaluation which is based on statistical measures. Precision evaluates the proportion of correctness for the sentences in the summary whereas recall is utilized to evaluate the proportion of relevant sentences included in the summary.

4.1 Precision

$$\text{Precision} = \frac{\{\text{Retrieved sentences}\} - \{\text{Relevant sentences}\}}{\{\text{Retrieved Sentences}\}}$$

The higher the precision value, the better is the efficiency of the system in reducing irrelevant Sentences

4.2 Recall

$$\text{Recall} = \frac{\{\text{Retrieved sentences}\} - \{\text{Relevant sentences}\}}{\{\text{relevant sentences}\}}$$

Higher the recall value, better the efficiency of the approach in selecting only relevant sentences.

4.3 F-Measure

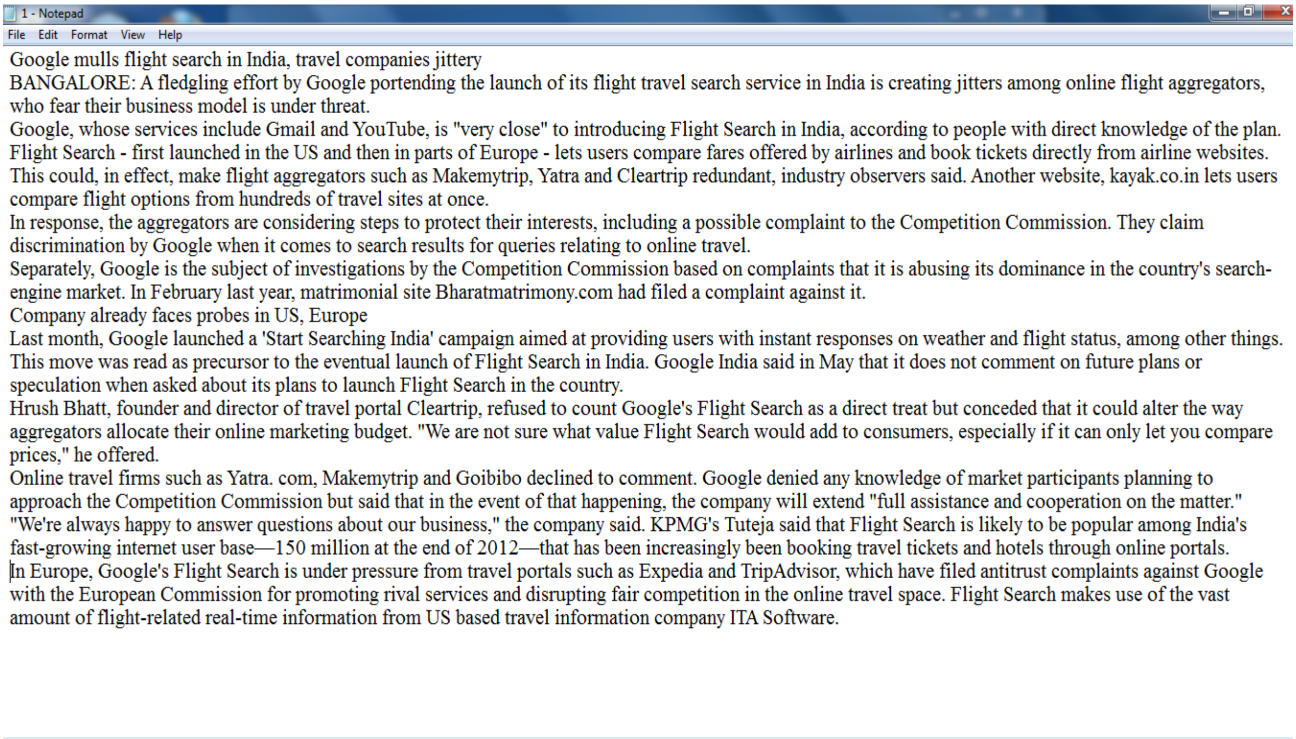
The weighted harmonic mean of precision and recall is called as F-measure

$$\text{F-measure} = \frac{2 \times \text{Precision} * \text{recall}}{\text{Precision} + \text{recall}}$$

5. EXPERIMENTAL RESULTS

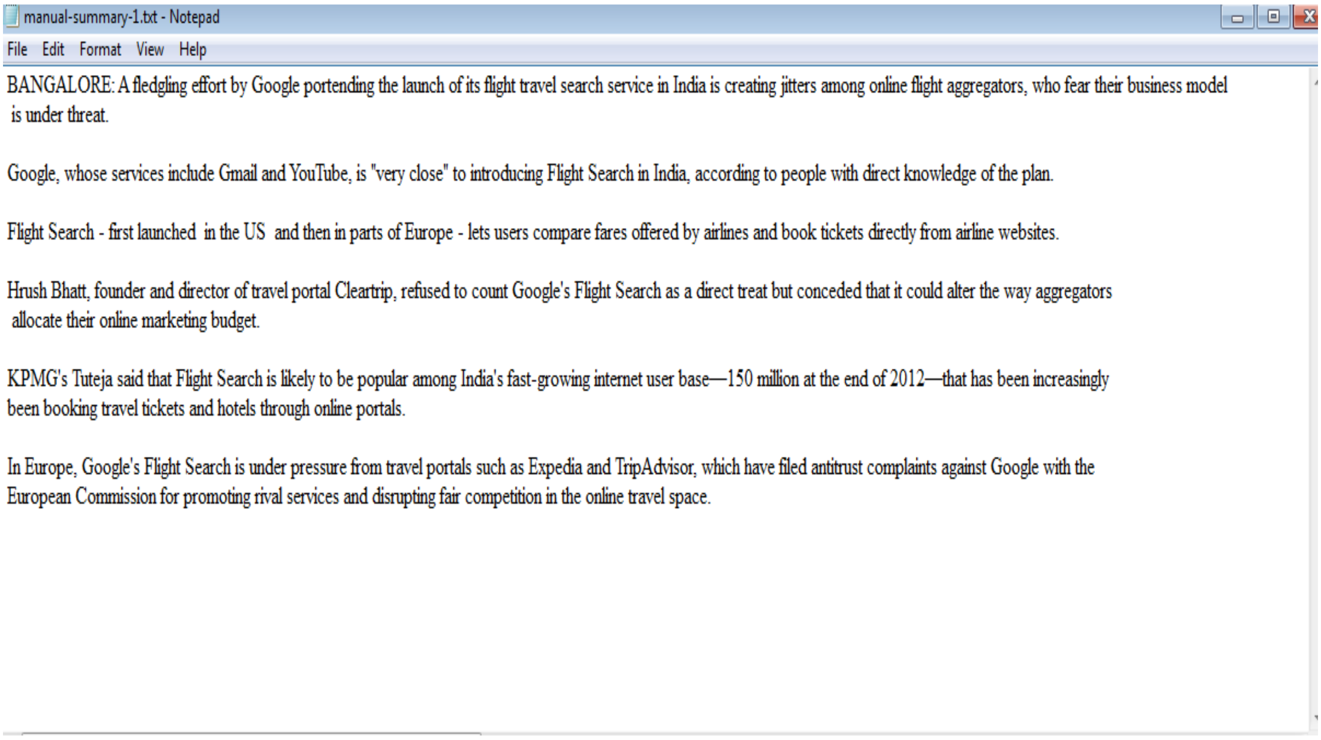
Three documents are taken in news domain. The original document, manually generated summaries and summaries generated by the above approach are shown below. The precision recall and F-measure are calculated for these three documents and they are compared with other two summarizers.

Original Document 1



Google mulls flight search in India, travel companies jittery
BANGALORE: A fledgling effort by Google portending the launch of its flight travel search service in India is creating jitters among online flight aggregators, who fear their business model is under threat.
Google, whose services include Gmail and YouTube, is "very close" to introducing Flight Search in India, according to people with direct knowledge of the plan.
Flight Search - first launched in the US and then in parts of Europe - lets users compare fares offered by airlines and book tickets directly from airline websites.
This could, in effect, make flight aggregators such as Makemytrip, Yatra and Cleartrip redundant, industry observers said. Another website, kayak.co.in lets users compare flight options from hundreds of travel sites at once.
In response, the aggregators are considering steps to protect their interests, including a possible complaint to the Competition Commission. They claim discrimination by Google when it comes to search results for queries relating to online travel.
Separately, Google is the subject of investigations by the Competition Commission based on complaints that it is abusing its dominance in the country's search-engine market. In February last year, matrimonial site Bharatmatrimony.com had filed a complaint against it.
Company already faces probes in US, Europe
Last month, Google launched a 'Start Searching India' campaign aimed at providing users with instant responses on weather and flight status, among other things. This move was read as precursor to the eventual launch of Flight Search in India. Google India said in May that it does not comment on future plans or speculation when asked about its plans to launch Flight Search in the country.
Hrush Bhatt, founder and director of travel portal Cleartrip, refused to count Google's Flight Search as a direct treat but conceded that it could alter the way aggregators allocate their online marketing budget. "We are not sure what value Flight Search would add to consumers, especially if it can only let you compare prices," he offered.
Online travel firms such as Yatra. com, Makemytrip and Goibibo declined to comment. Google denied any knowledge of market participants planning to approach the Competition Commission but said that in the event of that happening, the company will extend "full assistance and cooperation on the matter."
"We're always happy to answer questions about our business," the company said. KPMG's Tuteja said that Flight Search is likely to be popular among India's fast-growing internet user base—150 million at the end of 2012—that has been increasingly been booking travel tickets and hotels through online portals.
In Europe, Google's Flight Search is under pressure from travel portals such as Expedia and TripAdvisor, which have filed antitrust complaints against Google with the European Commission for promoting rival services and disrupting fair competition in the online travel space. Flight Search makes use of the vast amount of flight-related real-time information from US based travel information company ITA Software.

Ideal Summary of Document 1



BANGALORE: A fledgling effort by Google portending the launch of its flight travel search service in India is creating jitters among online flight aggregators, who fear their business model is under threat.

Google, whose services include Gmail and YouTube, is "very close" to introducing Flight Search in India, according to people with direct knowledge of the plan.

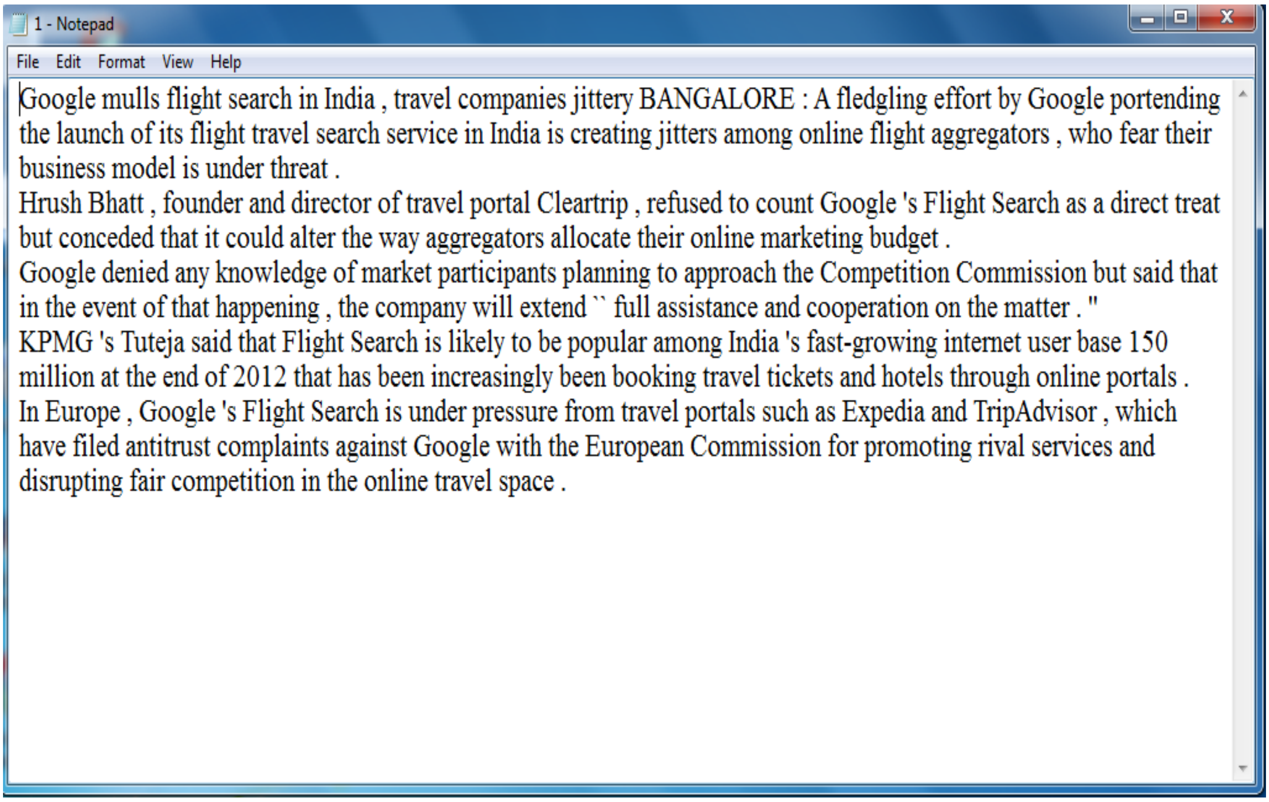
Flight Search - first launched in the US and then in parts of Europe - lets users compare fares offered by airlines and book tickets directly from airline websites.

Hrush Bhatt, founder and director of travel portal Cleartrip, refused to count Google's Flight Search as a direct treat but conceded that it could alter the way aggregators allocate their online marketing budget.

KPMG's Tuteja said that Flight Search is likely to be popular among India's fast-growing internet user base—150 million at the end of 2012—that has been increasingly been booking travel tickets and hotels through online portals.

In Europe, Google's Flight Search is under pressure from travel portals such as Expedia and TripAdvisor, which have filed antitrust complaints against Google with the European Commission for promoting rival services and disrupting fair competition in the online travel space.

Summary of Document 1 generated by our Summarizer



Google mulls flight search in India , travel companies jittery BANGALORE : A fledgling effort by Google portending the launch of its flight travel search service in India is creating jitters among online flight aggregators , who fear their business model is under threat .

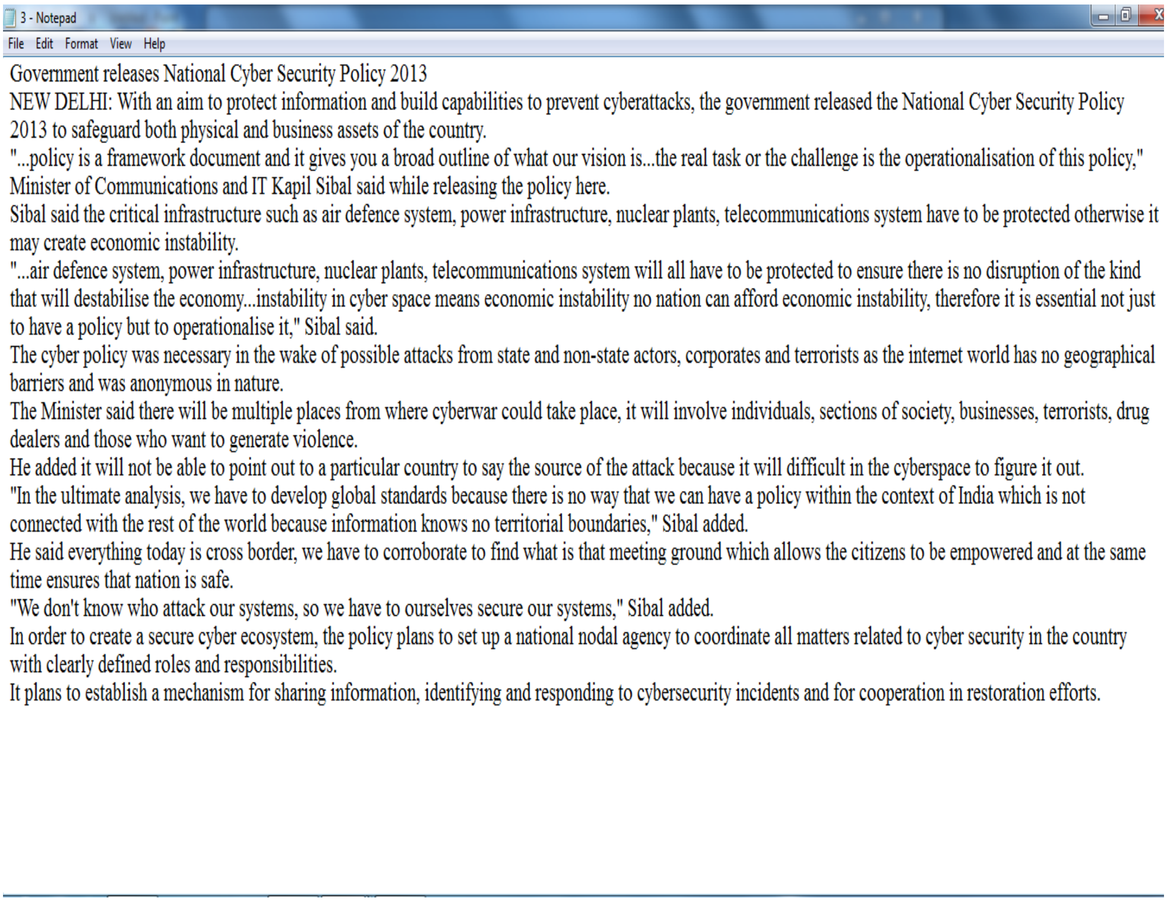
Hrush Bhatt , founder and director of travel portal Cleartrip , refused to count Google 's Flight Search as a direct treat but conceded that it could alter the way aggregators allocate their online marketing budget .

Google denied any knowledge of market participants planning to approach the Competition Commission but said that in the event of that happening , the company will extend `` full assistance and cooperation on the matter . "

KPMG 's Tuteja said that Flight Search is likely to be popular among India 's fast-growing internet user base 150 million at the end of 2012 that has been increasingly been booking travel tickets and hotels through online portals .

In Europe , Google 's Flight Search is under pressure from travel portals such as Expedia and TripAdvisor , which have filed antitrust complaints against Google with the European Commission for promoting rival services and disrupting fair competition in the online travel space .

Original Document 2



Government releases National Cyber Security Policy 2013

NEW DELHI: With an aim to protect information and build capabilities to prevent cyberattacks, the government released the National Cyber Security Policy 2013 to safeguard both physical and business assets of the country.

"...policy is a framework document and it gives you a broad outline of what our vision is...the real task or the challenge is the operationalisation of this policy," Minister of Communications and IT Kapil Sibal said while releasing the policy here.

Sibal said the critical infrastructure such as air defence system, power infrastructure, nuclear plants, telecommunications system have to be protected otherwise it may create economic instability.

"...air defence system, power infrastructure, nuclear plants, telecommunications system will all have to be protected to ensure there is no disruption of the kind that will destabilise the economy...instability in cyber space means economic instability no nation can afford economic instability, therefore it is essential not just to have a policy but to operationalise it," Sibal said.

The cyber policy was necessary in the wake of possible attacks from state and non-state actors, corporates and terrorists as the internet world has no geographical barriers and was anonymous in nature.

The Minister said there will be multiple places from where cyberwar could take place, it will involve individuals, sections of society, businesses, terrorists, drug dealers and those who want to generate violence.

He added it will not be able to point out to a particular country to say the source of the attack because it will difficult in the cyberspace to figure it out.

"In the ultimate analysis, we have to develop global standards because there is no way that we can have a policy within the context of India which is not connected with the rest of the world because information knows no territorial boundaries," Sibal added.

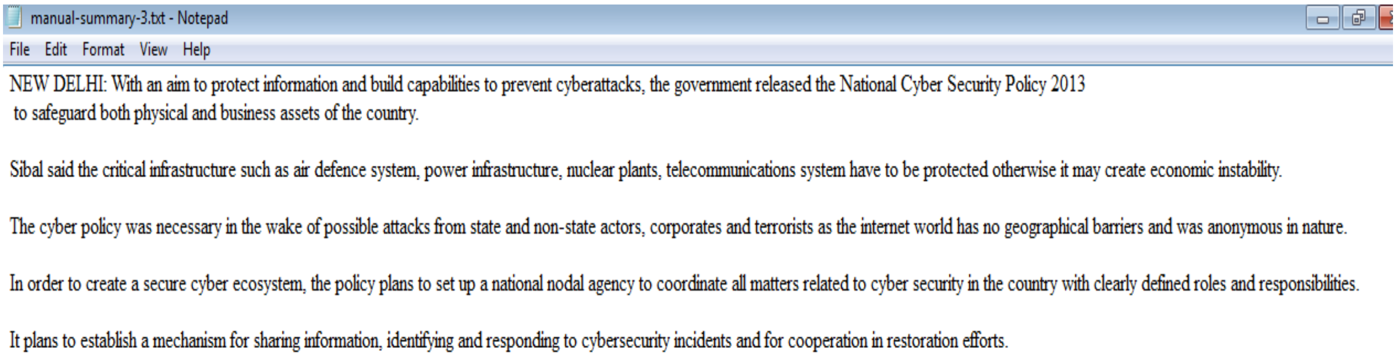
He said everything today is cross border, we have to corroborate to find what is that meeting ground which allows the citizens to be empowered and at the same time ensures that nation is safe.

"We don't know who attack our systems, so we have to ourselves secure our systems," Sibal added.

In order to create a secure cyber ecosystem, the policy plans to set up a national nodal agency to coordinate all matters related to cyber security in the country with clearly defined roles and responsibilities.

It plans to establish a mechanism for sharing information, identifying and responding to cybersecurity incidents and for cooperation in restoration efforts.

Ideal Summary of Document 2



NEW DELHI: With an aim to protect information and build capabilities to prevent cyberattacks, the government released the National Cyber Security Policy 2013 to safeguard both physical and business assets of the country.

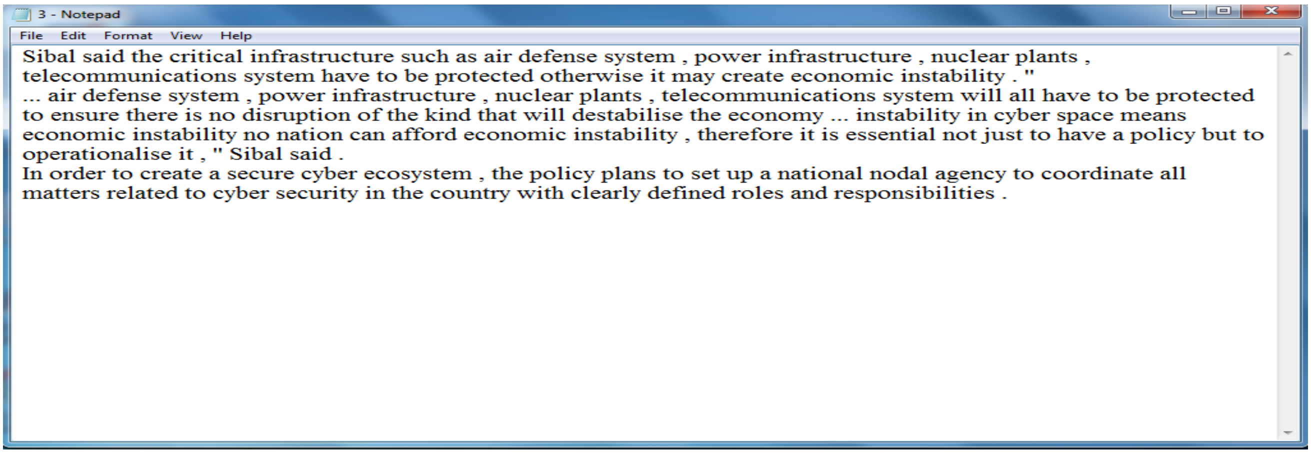
Sibal said the critical infrastructure such as air defence system, power infrastructure, nuclear plants, telecommunications system have to be protected otherwise it may create economic instability.

The cyber policy was necessary in the wake of possible attacks from state and non-state actors, corporates and terrorists as the internet world has no geographical barriers and was anonymous in nature.

In order to create a secure cyber ecosystem, the policy plans to set up a national nodal agency to coordinate all matters related to cyber security in the country with clearly defined roles and responsibilities.

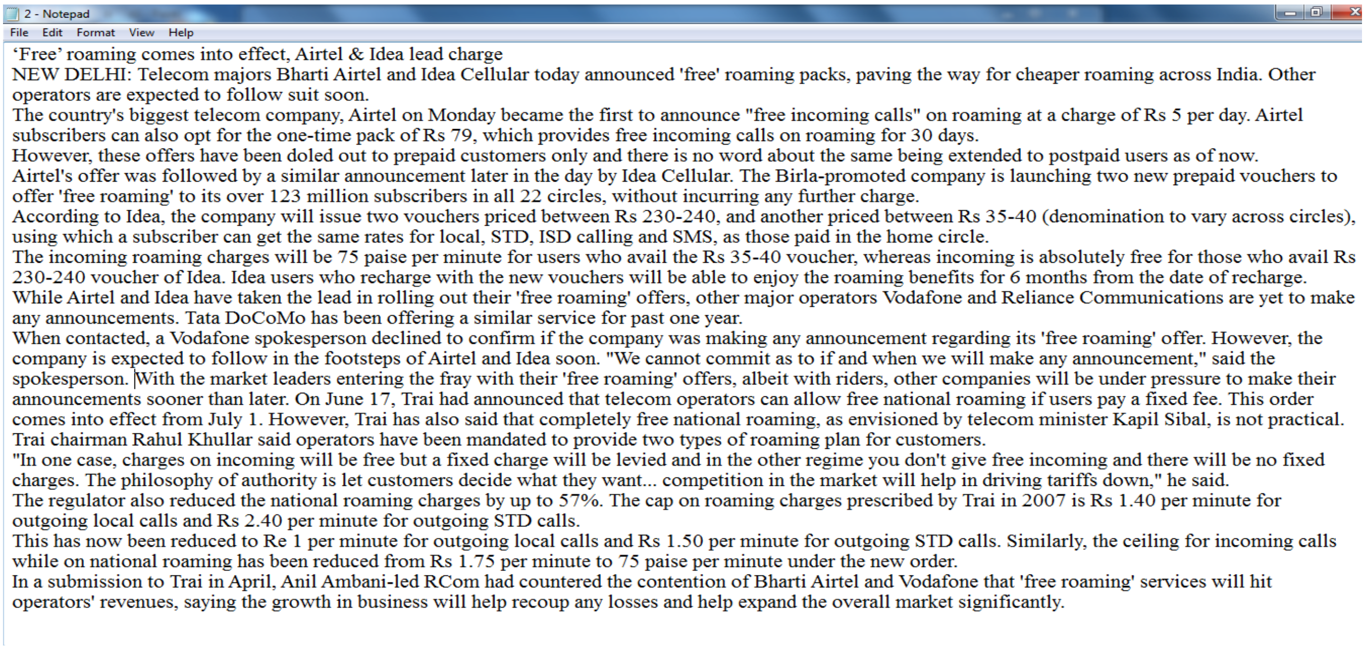
It plans to establish a mechanism for sharing information, identifying and responding to cybersecurity incidents and for cooperation in restoration efforts.

Summary of Document 2 generated by our summarizer



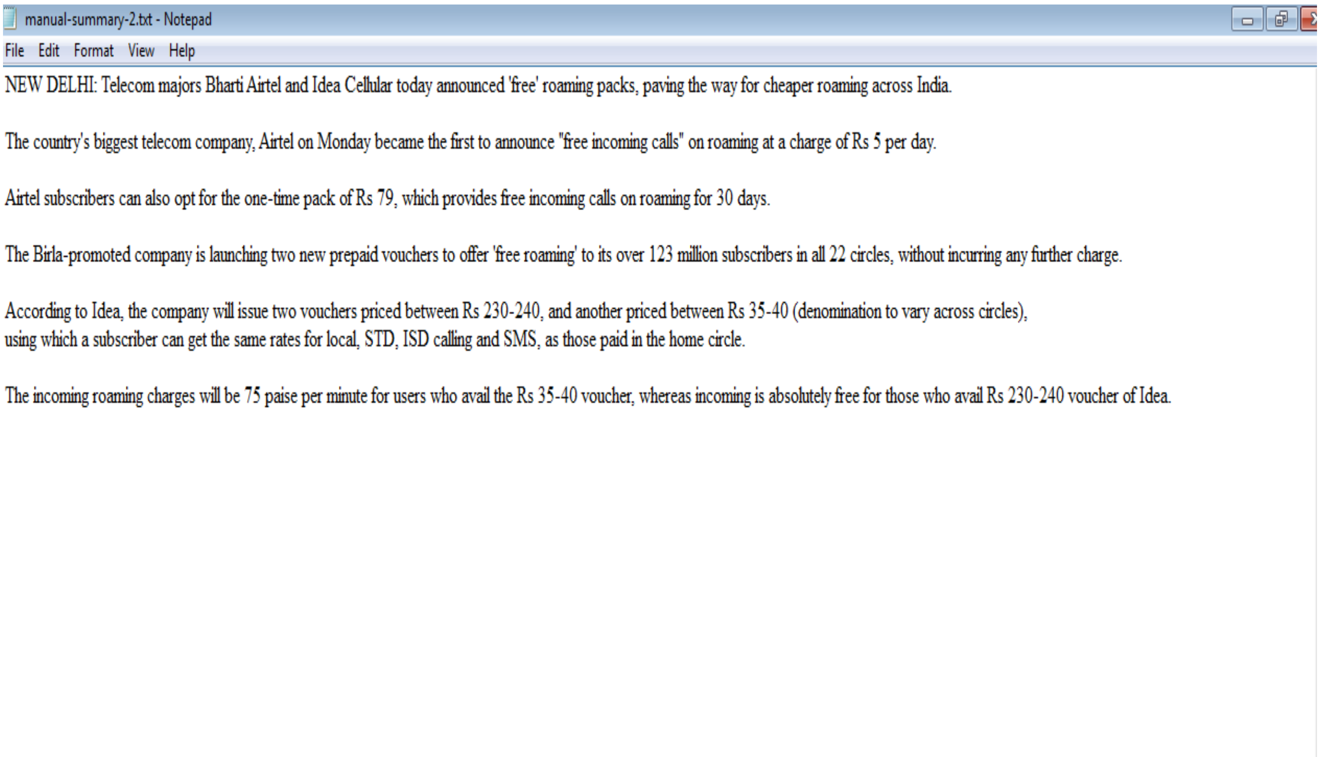
Sibal said the critical infrastructure such as air defense system , power infrastructure , nuclear plants , telecommunications system have to be protected otherwise it may create economic instability . "
... air defense system , power infrastructure , nuclear plants , telecommunications system will all have to be protected to ensure there is no disruption of the kind that will destabilise the economy ... instability in cyber space means economic instability no nation can afford economic instability , therefore it is essential not just to have a policy but to operationalise it , " Sibal said .
In order to create a secure cyber ecosystem , the policy plans to set up a national nodal agency to coordinate all matters related to cyber security in the country with clearly defined roles and responsibilities .

Original document 3



'Free' roaming comes into effect, Airtel & Idea lead charge
NEW DELHI: Telecom majors Bharti Airtel and Idea Cellular today announced 'free' roaming packs, paving the way for cheaper roaming across India. Other operators are expected to follow suit soon.
The country's biggest telecom company, Airtel on Monday became the first to announce "free incoming calls" on roaming at a charge of Rs 5 per day. Airtel subscribers can also opt for the one-time pack of Rs 79, which provides free incoming calls on roaming for 30 days.
However, these offers have been doled out to prepaid customers only and there is no word about the same being extended to postpaid users as of now.
Airtel's offer was followed by a similar announcement later in the day by Idea Cellular. The Birla-promoted company is launching two new prepaid vouchers to offer 'free roaming' to its over 123 million subscribers in all 22 circles, without incurring any further charge.
According to Idea, the company will issue two vouchers priced between Rs 230-240, and another priced between Rs 35-40 (denomination to vary across circles), using which a subscriber can get the same rates for local, STD, ISD calling and SMS, as those paid in the home circle.
The incoming roaming charges will be 75 paise per minute for users who avail the Rs 35-40 voucher, whereas incoming is absolutely free for those who avail Rs 230-240 voucher of Idea. Idea users who recharge with the new vouchers will be able to enjoy the roaming benefits for 6 months from the date of recharge.
While Airtel and Idea have taken the lead in rolling out their 'free roaming' offers, other major operators Vodafone and Reliance Communications are yet to make any announcements. Tata DoCoMo has been offering a similar service for past one year.
When contacted, a Vodafone spokesperson declined to confirm if the company was making any announcement regarding its 'free roaming' offer. However, the company is expected to follow in the footsteps of Airtel and Idea soon. "We cannot commit as to if and when we will make any announcement," said the spokesperson. With the market leaders entering the fray with their 'free roaming' offers, albeit with riders, other companies will be under pressure to make their announcements sooner than later. On June 17, Trai had announced that telecom operators can allow free national roaming if users pay a fixed fee. This order comes into effect from July 1. However, Trai has also said that completely free national roaming, as envisioned by telecom minister Kapil Sibal, is not practical.
Trai chairman Rahul Khullar said operators have been mandated to provide two types of roaming plan for customers.
"In one case, charges on incoming will be free but a fixed charge will be levied and in the other regime you don't give free incoming and there will be no fixed charges. The philosophy of authority is let customers decide what they want... competition in the market will help in driving tariffs down," he said.
The regulator also reduced the national roaming charges by up to 57%. The cap on roaming charges prescribed by Trai in 2007 is Rs 1.40 per minute for outgoing local calls and Rs 2.40 per minute for outgoing STD calls.
This has now been reduced to Re 1 per minute for outgoing local calls and Rs 1.50 per minute for outgoing STD calls. Similarly, the ceiling for incoming calls while on national roaming has been reduced from Rs 1.75 per minute to 75 paise per minute under the new order.
In a submission to Trai in April, Anil Ambani-led RCom had countered the contention of Bharti Airtel and Vodafone that 'free roaming' services will hit operators' revenues, saying the growth in business will help recoup any losses and help expand the overall market significantly.

Ideal summary of document 3



NEW DELHI: Telecom majors Bharti Airtel and Idea Cellular today announced 'free' roaming packs, paving the way for cheaper roaming across India.

The country's biggest telecom company, Airtel on Monday became the first to announce "free incoming calls" on roaming at a charge of Rs 5 per day.

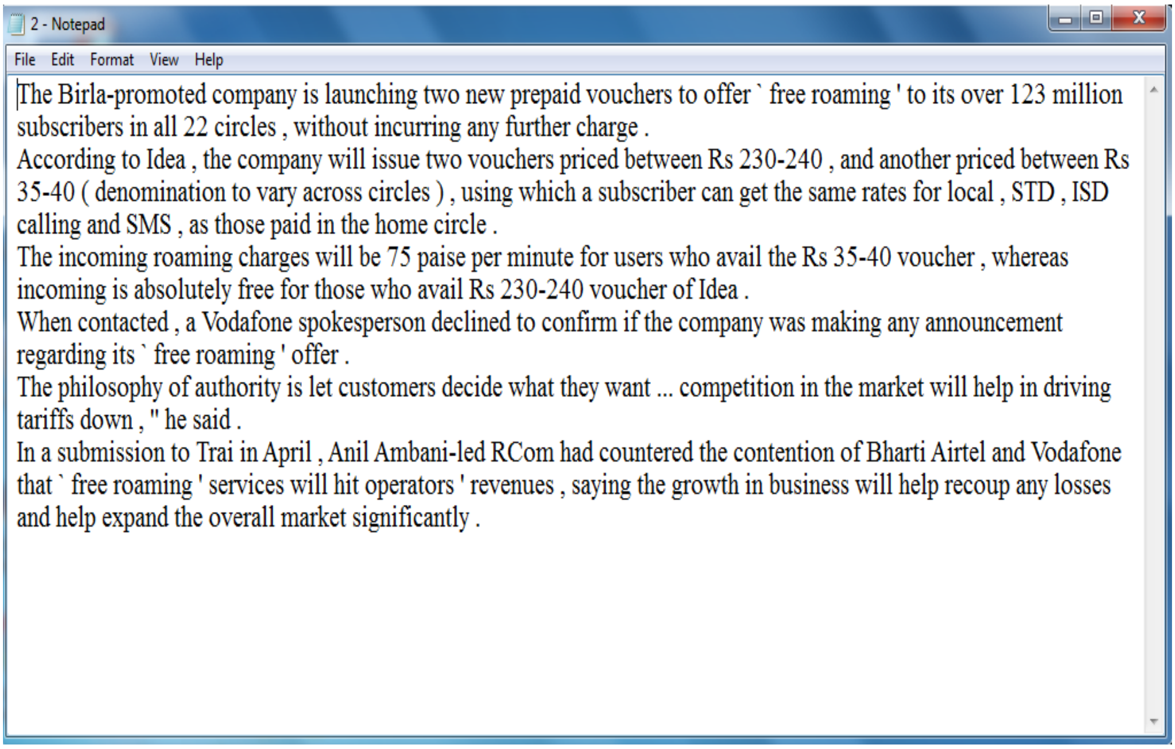
Airtel subscribers can also opt for the one-time pack of Rs 79, which provides free incoming calls on roaming for 30 days.

The Birla-promoted company is launching two new prepaid vouchers to offer 'free roaming' to its over 123 million subscribers in all 22 circles, without incurring any further charge.

According to Idea, the company will issue two vouchers priced between Rs 230-240, and another priced between Rs 35-40 (denomination to vary across circles),
using which a subscriber can get the same rates for local, STD, ISD calling and SMS, as those paid in the home circle.

The incoming roaming charges will be 75 paise per minute for users who avail the Rs 35-40 voucher, whereas incoming is absolutely free for those who avail Rs 230-240 voucher of Idea.

Generated Summary of Document 3

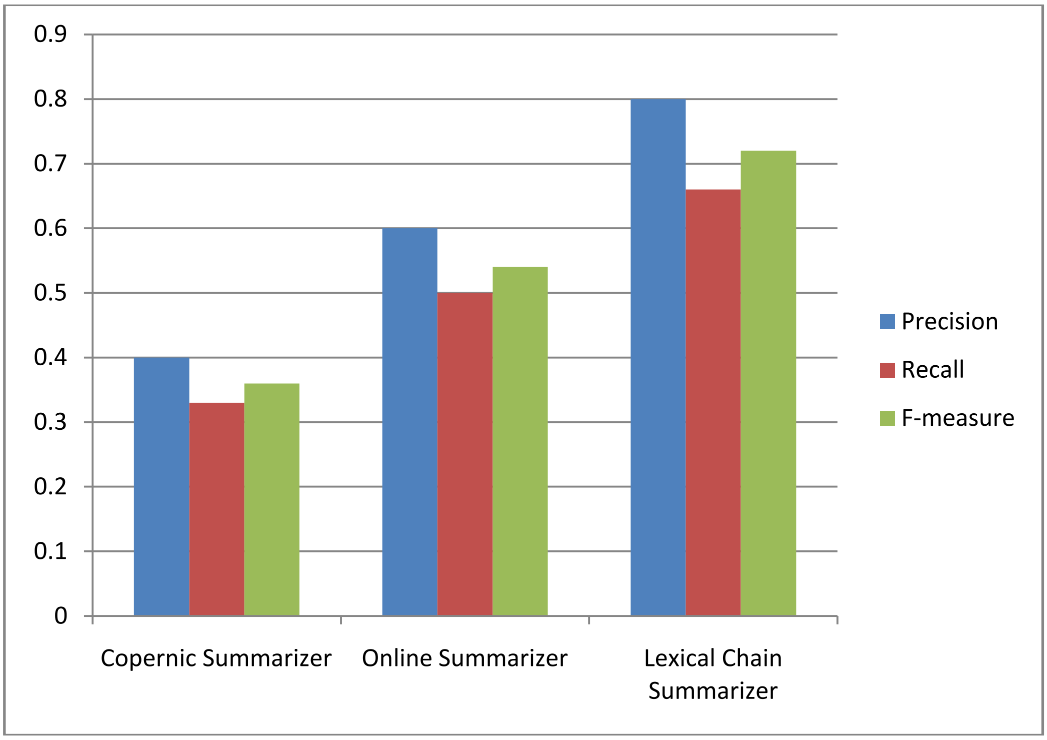


The Birla-promoted company is launching two new prepaid vouchers to offer ` free roaming ' to its over 123 million subscribers in all 22 circles , without incurring any further charge .
According to Idea , the company will issue two vouchers priced between Rs 230-240 , and another priced between Rs 35-40 (denomination to vary across circles) , using which a subscriber can get the same rates for local , STD , ISD calling and SMS , as those paid in the home circle .
The incoming roaming charges will be 75 paise per minute for users who avail the Rs 35-40 voucher , whereas incoming is absolutely free for those who avail Rs 230-240 voucher of Idea .
When contacted , a Vodafone spokesperson declined to confirm if the company was making any announcement regarding its ` free roaming ' offer .
The philosophy of authority is let customers decide what they want ... competition in the market will help in driving tariffs down , " he said .
In a submission to Trai in April , Anil Ambani-led RCom had countered the contention of Bharti Airtel and Vodafone that ` free roaming ' services will hit operators ' revenues , saying the growth in business will help recoup any losses and help expand the overall market significantly .

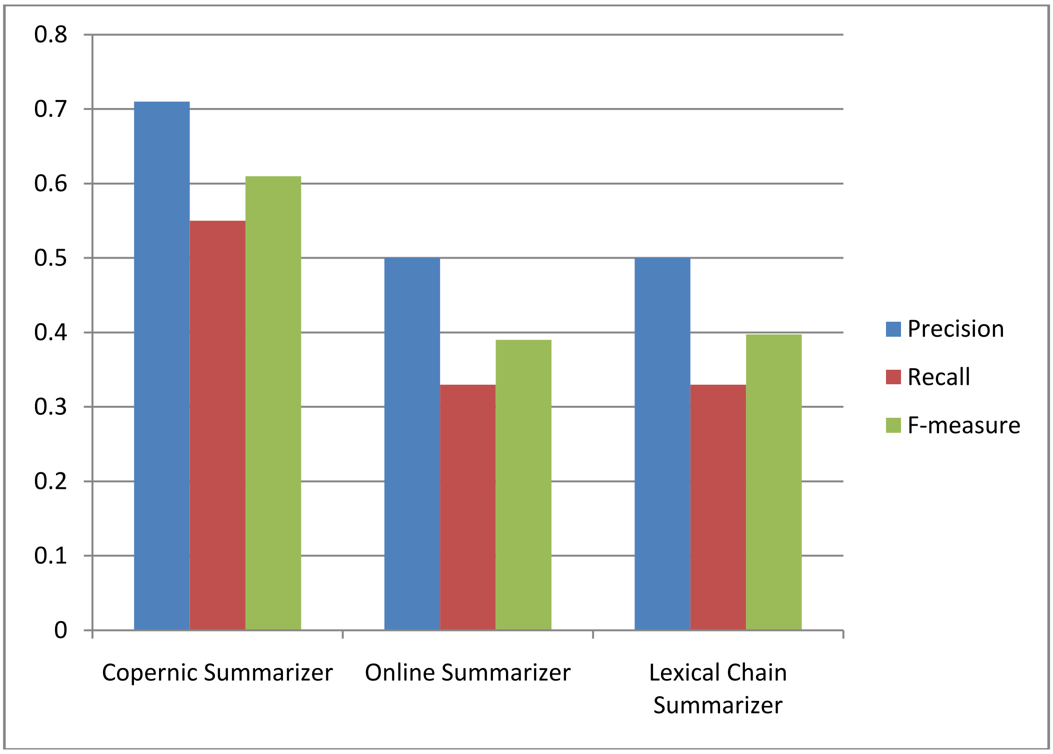
6. COMPARISION

This paper considers online summarizer from freesummarizer.com[7], Copernicus summarizer and our summarizer using lexical chains of sentences for comparison. The above three documents are used as input to all the three summarizers. The precision, recall and F-measure are used as performance measures for summary generated**.**

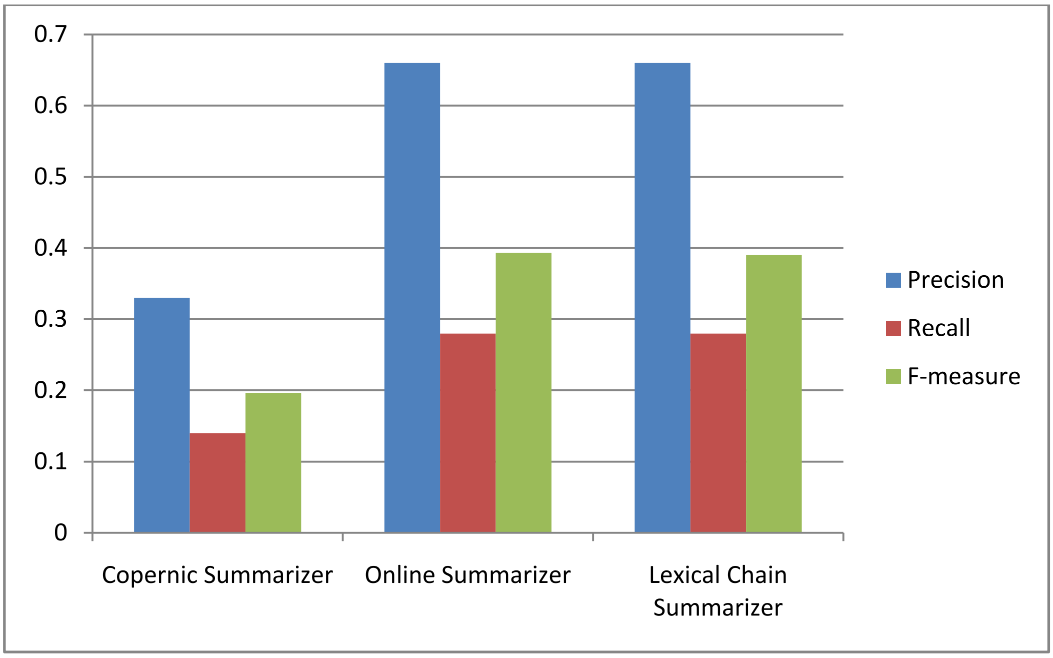
Document1



Document2:



Document3:



7. CONCLUSIONS

It is seen that for document 1 and document 3 our summarizer performs better than Copernicus summarizer. and online summarizer. For document 2, It performs equally as online summarizer but less efficient than Copernicus summarizer. Our summarizer is better as it also considers the semantic analysis of the document & correlation of sentences for generating the summary.

REFERENCES

[1] Canasai Kruengkari and Chuleer at Jaruskulchai, "Generic Text Summarization Using Local and Global Properties of Sentences", Proceedings of the IEEE/WIC international Conference on Web Intelligence (WI’03), 2003.

[2] Morris, J. and G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of the text. In Computational Linguistics, 18(1):pp21-45. 1991.

[3] Barzilay, Regina and Michael Elhadad. Using Lexical Chains for Text Summarization. in Proceedings of the Intelligent Scalable Text Summarization Workshop.(ISTS’97), ACL Madrid, 1997.

[4] Rene Arnulfo Garcia-Hera ndez and Yulia Ledeneva, “Word Sequence Models for Single Text Summarization”, IEEE,44-48, 2009.

[5] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, Jaime Carbonell, Summarizing text documents: Sentence Selection and Evaluation Metrics, Language Technologies Institute, Carnegie Mellon University.

[6] Khosrow Kaikhah, "Automatic Text summarization with Neural Networks", in Proceedings of second international Conference on intelligent systems, IEEE, 40-44, Texas, USA, June 2004.

[7] www.freesummarizer.com/summarize/