JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Issue-2 E-ISSN: 2347-2693

Automatic News Article Summarization

Laxmi B. Rananavare[1*], P. Venkata Subba Reddy[2]

[1*]Dept. of CSE, Sri Venkateswara University College of Engineering, Tirupathi, India
[2]Dept. of CSE, Sri Venkateswara University College of Engineering, Tirupathi, India

[*]*Corresponding Author: rananavare@yahoo.com, Tel.: +919880431355*

Available online at: www.ijcseonline.org

Received: 23/Jan/2018, Revised: 31/Jan/2018, Accepted: 15/Feb/2018, Published: 28/Feb/2018

Abstract: A summary condenses a lengthy document by highlighting salient features. It helps reader to understand completely just by reading summary so that the reader can save time and also can decide whether to go through the entire document. Summaries should be shorter than the original article so make sure that to select only pertinent information to include the article. The main goal of newspaper article summary is, the readers to walk away with knowledge on what the newspaper article is all about without the need to read the entire article. This work proposes a news article summarization system which access information from various local on-line newspapers automatically and summarizes information using heterogeneous articles. To make ad-hoc keyword based extraction of news articles, the system uses a tailor-made web crawler which crawls the websites for searching relevant articles. Computational Linguistic techniques mainly Triplet Extraction, Semantic Similarity calculation and OPTICS clustering with DBSCAN is used alongside a sentence selection heuristic to generate coherent and cogent summaries irrespective of the number of articles supplied to the engine. The performance evaluation is done using ROUGE metric.

Keywords:Text Summarization, Natural Language Processing, News Paper Articles, Intelligence mining, RDF Triplets ,NER.

I. INTRODUCTION

Now a days the large volume of information in electronic form is increasing rapidly. It can be structured data like databases, company legacy data; or unstructured data like text, images etc. About 85 and 90% of data is held in unstructured form [1]. Therefore, text mining is necessary for extracting and managing useful information from unstructured sets of data, such as news reports, emails and webpages, using a various text mining techniques. Hence, text mining has become an important and active research field. It is well known that text mining techniques have mostly been developed for the English language because most electronic data is in English. Using this to our advantage, it is an obvious next step to employ these techniques for sifting through the multitude of available on-line data to mine facts and figures from various sources and then summarize them efficiently to use in tracking various events in and around an area under Police jurisprudence. In this paper, the information extraction of news articles based on computational linguistic techniques to summarize the text. The summarization process involves filtering, highlighting and organizing information which is concise, coherent and faithful to the original document. The key tasks in summarization is as follows,

1. Automatically extract on-line articles from news websites based on a keyword.
2. Divide entire articles as a group of sentences, which acts as the dataset for further processing.
3. Representing sentences in a machine readable and understandable format.
4. Detecting semantic similarity between sentences so as to eliminate factual redundancy in summary.
5. Clustering similar sentences to distinguish between semantically different sentences.
6. Picking sentences amongst clusters which represent the information presented by the corresponding cluster.
7. Arranging the sentences chronologically to display the developments as they happened.

II. RELATED WORK

Extraction of a single summary from multiple documents has gained interest since mid-1990s, most applications being in the domain of news articles. Several Web based news clustering systems were inspired by research on multi-document summarization, for example Columbia NewsBlaster, or News In Essence. This is different from single-document summarization since the problem involves multiple sources of information that overlap and supplement each other, and removal of redundant facts which are presented in a semantically similar but grammatically different structure. The key factor in multi document summarization is to recognize the novelty and ensure that the final summary is both coherent and complete.

Various approaches in Multi-Document Summarization are as follows:

1. Abstraction and Information Fusion: The summaries are created by merging facts from various document sources to generate an informational summary of the same. These techniques also employed the use of a linguistic generator to create sentences out of words selected based on statistical analysis techniques like TF-IDF scores, noun pronoun and verb weights etc. [2].

2. Topic Driven Summarization: The summary consists of set of topic-related documents which relevant to the application's or user's need.. This can be done by employing weighted keyword analysis [3], topic signatures [4] and Statistical methods like Latent Dirichlet Allocation [5] Latent Semantic Indexing and Probabilistic Latent Semantic Analysis [6].

3. Graph Based Summarization: Graph and ontology based methods usually use fuzzy logic to determine which of the data is relevant to each other to avoid redundancy in summarization or by supervised learning approaches by guiding the system to learn how to select the correct sentences for summarization [7]. Using classifiers sentences are also picked from the document Semantic Graph. A document is represented as a graph and each node represents the occurrence of a single word (i.e., one word together with its position in the text) [8].

4. Centroid Based Summarization: These techniques use clustering of sentences and then using centroids of said clusters to generate informative summaries [9]. These techniques do not employ a language generation module, thus making it easy to scale and remain domain independent.

- **ADVANTAGES AND DISADVANTAGES**

ADVANTAGES :

1. Multi-document summarization generates summary that are concise, coherent and non-redundant. With different opinions being put together and outlined, every topic is described from multiple perspectives within a single document.

2. Automatic summaries present information extracted from multiple sources algorithmically, without any editorial touch or subjective human intervention, thus making it completely unbiased.

3. The large amount of data available about a topic is concisely presented and thus makes it easier to study and remain informed.

DISADVANTAGES:

1. The need to eliminate redundancy of data.

2. Sometimes multiple facts are contradictory (death toll, time and date etc.)

III. METHODOLOGY

Devising an application which automatically collects digital on-line news articles based on a key-word from local newspaper articles and then summarizes those using Natural Language Processing techniques to semantically cluster sentences and the extract sentences from said clusters using Centroid-Based sentence summarization techniques.

WORK OBJECTIVES

1. Collecting data for building a corpora of newspaper articles.
2. Develop a mechanism for breaking down actual sentences into a representative format for semantic analysis.
3. Clustering semantically similar sentences to derive non-redundant summaries of the same.
4. Evaluating the summaries generated by the system.

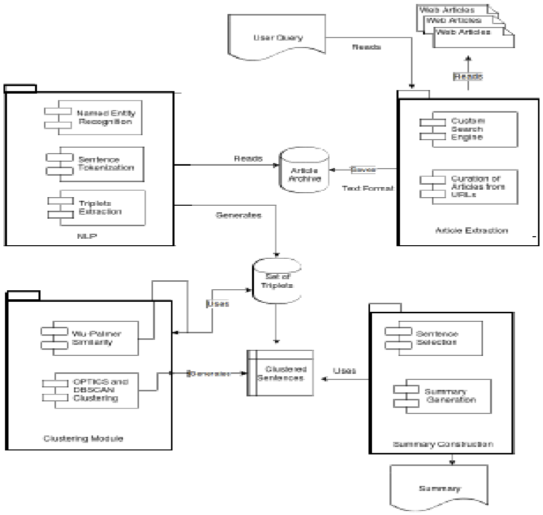


Figure 1: Application work flow

The user will be first asked to enter a query regarding which the summary has to be generated. The user the selects the newspapers he/she wants to collect the articles from. The articles are the extracted by web crawling and scraping and then saved in the system as archives and as the dataset for summarization. This Dataset is the further divided into individual sentences and the triplet extraction algorithm is run on each sentence and the result is saved for creating of the similarity matrix. Once the similarity matrix is created, a clustering algorithm is applied on it and get a cluster of all similar sentences. A sentence is selected from each cluster and then put in the summary based on heuristics which shall be further discussed in following sections. Once the summary is created, it is displayed and the user also has an option of picking the sentences which are better suited for their needs as this information can then be used to tailor a summary engine which uses supervised machine learning techniques and thus gives better results.

There are various steps involved in the process of summarization which can be briefly outlined as follows:

- Breaking down all the articles into individual sentences.
- Breaking down individual sentences into the RDF Triplet format or the Subject Verb Object triplet format for semantic analysis.
- Named Entity Recognition and Stemming.
- Calculating semantic similarities between triplets.
- Clustering Triplets from Similarity Matrix
- Sentence Selection

Each of these steps will be explained below,

A. SENTENCE TOKENIZATION

Sentence tokenization refers to the practice of dividing a text into a group of sentences. A news article as a whole is basically a collection of interrelated sentences. Since the structure of a news article is usually rigid and uniform [10], it is easier and computationally efficient to parse through the article sentence-wise instead of treating it as one single entity. It is both memory as well as time consuming to semantically analyze the entire article and hence it was decided to treat the entire event which is queried by the user to be treated to have a group of sentences comprising all news articles as the base dataset. The task then became to summarize the article information from the group of sentences rather than per article basis.

It is difficult to understand the semantics automatically from an entire article and hence it is necessary to break down the article into a set of sentences. This is done by the using the Punkt sentences tokenizer in NLTK tool-kit. It is an implementation of the Unsupervised Multilingual Sentence Boundary Detection Algorithm designed in [11].

They proposed to approach sentence boundary detection by first determining possible abbreviations in the text. They do so by identifying three major characteristics of abbreviations.

- An abbreviation is rather compact i.e. there is a close bond between a period and the letter preceding it.
- Abbreviations tend to be short.
- Experimental characterization of internal periods in abbreviations.

Using such heuristics, they built a classifier which determined whether a period was after the end of a sentence or followed preceded by and abbreviation, initial or an ordinal with 99.2% accuracy. Using this model, divide the article into a group of sentences which then acts as the base dataset to glean information about the article.

B. TRIPLET EXTRACTION

A Triplet consists of subject and object, the relation being the predicate of a given sentence. The aim here is to extract sets of the form subject, predicate, object out of syntactically parsed sentences. Basically a triplet is used to give an exact semantic sense of what a sentence is talking about. Instead of using the whole sentence to derive meaning; a triplet just uses three words to determine what the sentence is talking about.

To begin with; the sentence is first parsed to understand it's grammar by using the Stanford Treebank Parser. Stanford Parser is a natural language parser developed by Dan Klein and Christopher Manning from The Stanford NLP Group [12]. The package contains a Java implementation of the Treebank parser; a graphical user interface is also available, for parse tree visualization called *Stanford Tregex*. A treebank is a text corpus where each sentence belonging to the corpus has a syntactic structure added to it. In a treebank parser, a sentence (S) is represented by the parser as a tree having three children: a noun phrase (NP), a verbal phrase (VP) and the period (.). The root of the tree will be S. Triplet Extraction is done as follows:

- To find the subject of the sentence and apply a Breadth First Search in the NP sub-tree and select the first descendant of NP that is a noun.
- To find the predicate of the sentence, search for the deepest verb descendant in VP and assign that as the predicate.
- To find objects search in three different sub-trees. The sub-trees are: PP (prepositional phrase), NP and ADJP (adjective phrase). In NP and PP search for the first noun, while in ADJP find the first adjective.

```
Algorithm 1: Triplet Extraction
Data: sentence

Result: A solution or a failure

result  EXTRACT← SUBJECT(NP subtree) U

        EXTRACT PREDICATE(VP subtree) U

        EXTRACT OBJECT(VP subtree)

if result $\neq$ failure then

          return result

else
```

```
    return failure
end
```

```
Algorithm 2: EXTRACT ATTRIBUTES
Data: word
Result: A solution or a failure
/* search among the word's siblings */
if adjective(word) then
    result   all RB siblings
else
    if noun(word) then
        result ← all JJ, ADJP, NP siblings
    else
        if verb(word) then
            result ←  all ADVP siblings
        end
    end
end
/* search among word's immediate ancestor siblings */
if noun(word) OR adjective(word) then
    if uncle = PP then
        result ← uncle subtree
    end
    else
        if verb(word) AND (uncle = verb) then
            result ← uncle subtree
        end
    end
if result ≠ failure then
    return result
else
    return failure
end
```

```
Algorithm 3: EXTRACT SUBJECT
Data: NP subtree
Result: A solution or a failure
subject ← first noun found in NP subtree;
subjectAttributes←EXTRACT ATTRIBUTES(subject);
result   subject U subjectAttributes;
if result ≠ failure then
    return result
else
    return failure
end
```

```
Algorithm 4: EXTRACT PREDICATE
Data: VP subtree
Result: A solution or a failure
predicate ← deepest verb found in VP subtree;
predicateAttributes← EXTRACT ATTRIBUTES(predicate);
result ← predicate U predicateAttributes;
if result ≠ failure then
    return result
else
    return failure
end
```

```
Algorithm 5: EXTRACT OBJECT
Data: VP subtree
Result: A solution or a failure
For each value in siblings do
    if value = NP or PP then
    object ← first noun in value
  else
    object ← first adjective in value;
    objectAttributes←EXTRACTATTRIBUTES(object)
  end
end
result ← object U objectAttributes;
if result ≠ failure then
    return result
else
    return failure
end
```

C. NAMED ENTITY RECOGNITION AND STEMMING

Named entities are atomic elements in the text such as persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. [13, 14] Named entity recognition (NER) is the task of identifying such named entities. In the 1990s, the NER concept was introduced at the Message Understanding Conferences (MUCs) to encourage the development of new and better methods of information extraction.

Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form; generally a written word form. A stemmer for English, for example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root cat, and "stems", "stemming", "stemmed" as based on stem. A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the base word, fish.

After triplet extraction, the verbs are still in their `ing'(infinitive) format or plurals which causes issues in semantic analysis as the NLP tools only recognize root words and thus there is need to stem the words before performing semantic analysis. For grammatical reasons, documents are going to use different forms of a word, such as organize, organization, orgarnizes and organizing. Additionally, there are families of derivationally related words with similar meanings, such as democracy, democratic, and democratization. In many situations, it seems as if it would be useful for a search for one of these words to return documents that contain another word in the set. For instance:

am, are, is categorized as 'be'

car, cars, car's, cars categorized as 'car'

Thus a sentence which was originally \Police arrived at the scene" Will be ultimately fed to the semantic analyser as <Police, come, scene >where arrived has been stemmed and lemmatized to the verb `come'.

IV. RESULTS AND DISCUSSION

It was observed that sometimes the news articles extracted were not always pertaining to the query submitted by the user and the problem was detected to be the way the google custom search engine operated in which even if part of the query was fulfilled by the article, it was given as an output[15]. And hence, once an article was extracted, a further processing layer was added which then searched for the instances of the keywords in the article and only of the majority of the keywords were present, it then saved the article for summarization or else the article was discarded.

RESULTS OBTAINED AFTER TRIPLETS AND NER

Consider an actual news article from the DUC 2001 dataset. The article is about a Hurricane called Andrew and it is as follows:

> SQUADS of workers fanned out across storm-battered Louisiana yesterday to begin a massive rebuilding effort after Hurricane Andrew had attended whole districts, killing two people and injuring dozens more, agencies report from Florida and New Orleans. However, local officials in Florida, hit earlier in the week by the hurricane, were critical of what they called a delay in supplying food, drinking water and other supplies for thousands of people in need. Federal emergency officials acknowledged distribution problems, Transportation Secretary Andrew Card yesterday promised 'dramatic' improvements within 24 hours and President George Bush last night ordered troops to Florida, without specifying a number. The government estimated it would cost Dollars 20bn-Dollars 30bn to tidy and rebuild in Florida, and to care for residents displaced by the storm. Louisiana state officials said they had no overall count of storm-related injuries but initial estimates reckoned fewer than 100. The Federal Emergency Management Agency said it was setting aside Dollars 77m to help Louisiana recover. Most of the storm's fury was spent against sparsely populated farming communities and swampland in the state, sparing it the widespread destruction caused in Florida, where 15 people died. Official estimates in Miami reported that the hurricane had wiped out the homes of one Dade County resident in eight - a quarter of a million people. Andrew had become little more than a strong rainstorm early yesterday, moving across Mississippi State and heading for the north-eastern US. Several of Louisiana's main industries were affected, including those of oysters and alligators.
>
> Wildlife and fisheries secretary Joe Herring estimated a 50 per cent decline in the alligator industry. The cotton and sugar-cane crops were threatened, the state agriculture department said.

When this article is given for processing the following triplets shown in table 1.

Triplets with NER				
Sentence Number	Subject	Predicate	Object	NER resolutions
1	SQUADS	fan	Louisiana	(Location, Louisiana)
2	local	hit	Hurricane	none
3	official	acknowledge	problem	none
4	government	estimate	cost	none
5	Louisiana	said	count	(Location, Louisiana)
6	federal	set	dollars	none
7	storm	spent	Fury	none
8	Miami	Wipe	home	(Location, Miami)
9	Andrew	Move	Mississippi	(Location, Mississippi) (Person, Andrew)
10	industries	Affect	oysters	none
11	Wildlife	Estimate	decline	none
12	cotton	Threatened	no object	none

Table 1: Some Triplets extracted from the article

Similarly triplets are calculated for all the sentences of each article and then these triplets are then given to the clustering engine for determining semantic similarity.

```
Algorithm 6: DBSCAN
Data: D, ϵ
Result: all points in a point P's ϵ-neighbourhood
For each Point P in Dataset D do
    if P is Visited then
        continue to next point
    end
    mark P as visited;
    if sizeof(NeighbourPts) <MinPts then
        mark P as NOISE
    else
        C = next cluster;
        expandCluster (P, NeighbourPts, C, ϵ, MinPts)
    end
end
Algorithm 7: expandCluster
Data: P, NeighbourPts, C, ϵ, MinPts
add P to cluster C;
for each point P' in NeighbourPts do
    if P' is not visited then
        mark P' as visited;
        NeighbourPts' = regionQuery(P', ϵ);
    end
```

```
        if sizeof(NeighbourPts') ≤ MinPts then
                NeighbourPts = NeighbourPts joined with
                NeighbourPts'
end
        if P' is not yet member of any cluster then
                add P' to cluster C
        end
end
```

Algorithm 8: regionQuery

Data: P, ϵ

Result: all points in a point P's ϵ-neighbourhood

CLUSTERS FROM THE EXAMPLE ARTICLE

Cluster Number	Triplets
1	<Squads,fan,Louisiana> <Agency,help,Louisiana>
2	<industries,a_ect,oysters > <_sheries,decline,percent>
3	<Andrew,move,Mississippi>

Table 2: 3 of the 7 clusters formed by all the triplets

SENTENCE SELECTION

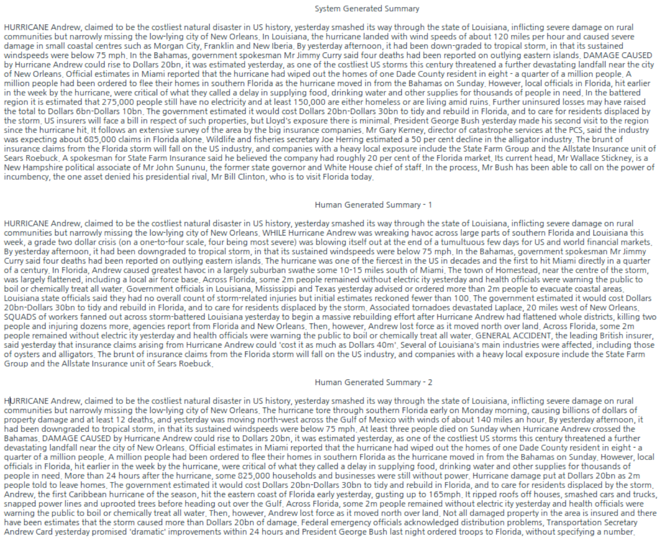
Once it is done clustering the sentences based on the information they provide and their semantic similarity pick a single sentence from each cluster which clearly represents information given by said cluster. This process is done as follows:

1. Since OPTICS already determines the clustering order, the sentences are as per centroids of each clusters, thus they give the most amount of information as to what the cluster pertains to.
2. Arrange the centroid sentences in a chronological manner with respect to date of publishing. This ensures factual chronology.
3. Output all the selected and sorted sentences as a wholesome summary.

DATASET

The dataset used for evaluation of the summary engine is called the DUC 2001 dataset. The DARPA program offered the opportunity to tackle summarization evaluation once again and a long-term road-map to guide this evaluation was created. This road-map provided guidance for the Document Understanding Conference (DUC), with a pilot run in 2000, and the first major evaluation in 2001. This model was evaluated on 3 topics mainly pertaining to political unrest, crime and natural disasters. Thus a total of 30 articles were used to generate three generic multi-document summaries and they were compared against 6 "gold standard" i.e. NIST generated summaries (2 for each topic) and were evaluated using the techniques.

COMPARISON OF SUMMARIES



System Generated Summary

HURRICANE Andrew, claimed to be the costliest natural disaster in US history, yesterday smashed its way through the state of Louisiana, inflicting severe damage on rural communities but narrowly missing the low-lying city of New Orleans. In Louisiana, the hurricane landed with wind speeds of about 120 miles per hour and caused severe damage in small coastal centres such as Morgan City, Franklin and New Iberia. By yesterday afternoon, it had been down-graded to tropical storm, in that its sustained windspeeds were below 75 mph. In the Bahamas, government spokesman Mr Jimmy Curry said four deaths had been reported on outlying eastern islands. DAMAGE CAUSED by Hurricane Andrew could rise to Dollars 20bn, it was estimated yesterday, as one of the costliest US storms this century threatened a further devastating landfall near the city of New Orleans. Official estimates in Miami reported that the hurricane had wiped out the homes of one Dade County resident in eight - a quarter of a million people. A million people had been ordered to flee their homes in southern Florida as the hurricane moved in from the Bahamas on Sunday. However, local officials in Florida, hit earlier in the week by the hurricane, were critical of what they called a delay in supplying food, drinking water and other supplies for thousands of people in need. In the battered region it is estimated that 275,000 people still have no electricity and at least 150,000 are either homeless or are living amid ruins. Further uninsured losses may have raised the total to Dollars 6bn-Dollars 10bn. The government estimated it would cost Dollars 20bn-Dollars 30bn to tidy and rebuild in Florida, and to care for residents displaced by the storm. US insurers will face a bill in respect of such properties, but Lloyd's exposure there is minimal. President George Bush yesterday made his second visit to the region since the hurricane hit. It follows an extensive survey of the area by the big insurance companies. Mr Gary Kerney, director of catastrophe services at the PCS, said the industry was expecting about 685,000 claims in Florida alone. Wildlife and fisheries secretary Joe Herring estimated a 50 per cent decline in the alligator industry. The brunt of insurance claims from the Florida storm will fall on the US industry, and companies with a heavy local exposure include the State Farm Group and the Allstate Insurance unit of Sears Roebuck. A spokesman for State Farm Insurance said he believed the company had roughly 20 per cent of the Florida market. Its current head, Mr Wallace Stickney, is a New Hampshire political associate of Mr John Sununu, the former state governor and White House chief of staff. In the process, Mr Bush has been able to call on the power of incumbency, the one asset denied his presidential rival, Mr Bill Clinton, who is to visit Florida today.

Human Generated Summary - 1

HURRICANE Andrew, claimed to be the costliest natural disaster in US history, yesterday smashed its way through the state of Louisiana, inflicting severe damage on rural communities but narrowly missing the low-lying city of New Orleans. WHILE Hurricane Andrew was wreaking havoc across large parts of southern Florida and Louisiana this week, a grade two dollar crisis (on a one-to-four scale, four being most severe) was blowing itself out at the end of a tumultuous few days for US and world financial markets. By yesterday afternoon, it had been downgraded to tropical storm, in that its sustained windspeeds were below 75 mph. In the Bahamas, government spokesman Mr Jimmy Curry said four deaths had been reported on outlying eastern islands. The hurricane was one of the fiercest in the US in decades and the first to hit Miami directly in a quarter of a century. In Florida, Andrew caused greatest havoc in a largely suburban swathe some 10-15 miles south of Miami. The town of Homestead, near the centre of the storm, was largely flattened, including a local air force base. Across Florida, some 2m people remained without electric ity yesterday and health officials were warning the public to boil or chemically treat all water. Government officials in Louisiana, Mississippi and Texas yesterday advised or ordered more than 2m people to evacuate coastal areas. Louisiana state officials said they had no overall count of storm-related injuries but initial estimates reckoned fewer than 100. The government estimated it would cost Dollars 20bn-Dollars 30bn to tidy and rebuild in Florida, and to care for residents displaced by the storm. Associated tornadoes devastated Laplace, 20 miles west of New Orleans. SQUADS of workers fanned out across storm-battered Louisiana yesterday to begin a massive rebuilding effort after Hurricane Andrew had flattened whole districts, killing two people and injuring dozens more, agencies report from Florida and New Orleans. Then, however, Andrew lost force as it moved north over land. Across Florida, some 2m people remained without electric ity yesterday and health officials were warning the public to boil or chemically treat all water. GENERAL ACCIDENT, the leading British insurer, said yesterday that insurance claims arising from Hurricane Andrew could 'cost it as much as Dollars 40m'. Several of Louisiana's main industries were affected, including those of oysters and alligators. The brunt of insurance claims from the Florida storm will fall on the US industry, and companies with a heavy local exposure include the State Farm Group and the Allstate Insurance unit of Sears Roebuck.

Human Generated Summary - 2

HURRICANE Andrew, claimed to be the costliest natural disaster in US history, yesterday smashed its way through the state of Louisiana, inflicting severe damage on rural communities but narrowly missing the low-lying city of New Orleans. The hurricane tore through southern Florida early on Monday morning, causing billions of dollars of property damage and at least 12 deaths, and yesterday was moving north-west across the Gulf of Mexico with winds of about 140 miles an hour. By yesterday afternoon, it had been downgraded to tropical storm, in that its sustained windspeeds were below 75 mph. At least three people died on Sunday when Hurricane Andrew crossed the Bahamas. DAMAGE CAUSED by Hurricane Andrew could rise to Dollars 20bn, it was estimated yesterday, as one of the costliest US storms this century threatened a further devastating landfall near the city of New Orleans. Official estimates in Miami reported that the hurricane had wiped out the homes of one Dade County resident in eight - a quarter of a million people. A million people had been ordered to flee their homes in southern Florida as the hurricane moved in from the Bahamas on Sunday. However, local officials in Florida, hit earlier in the week by the hurricane, were critical of what they called a delay in supplying food, drinking water and other supplies for thousands of people in need. More than 24 hours after the hurricane, some 825,000 households and businesses were still without power. Hurricane damage put at Dollars 20bn as 2m people told to leave homes. The government estimated it would cost Dollars 20bn-Dollars 30bn to tidy and rebuild in Florida, and to care for residents displaced by the storm. Andrew, the first Caribbean hurricane of the season, hit the eastern coast of Florida early yesterday, gusting up to 165mph. It ripped roofs off houses, smashed cars and trucks, snapped power lines and uprooted trees before heading out over the Gulf. Across Florida, some 2m people remained without electric ity yesterday and health officials were warning the public to boil or chemically treat all water. Then, however, Andrew lost force as it moved north over land. Not all damaged property in the area is insured and there have been estimates that the storm caused more than Dollars 20bn of damage. Federal emergency officials acknowledged distribution problems, Transportation Secretary Andrew Card yesterday promised 'dramatic' improvements within 24 hours and President George Bush last night ordered troops to Florida, without specifying a number.

Fig 2: Summary Comparison

The main evaluation metrics are precision, recall and F-score. Precision (P) is the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the system summary. Recall (R) is the number of sentences occurring in both system and ideal summaries divided by the number of sentences in the ideal summary. F-score is a composite measure that combines precision and recall. The basic way how to compute the F-score is to count a harmonic average of precision and recall:

F-score = (2*P*R) / (P+R)

With respect to the F-score computations, here are the results,

Total sentences in System Summary = 20

Total Sentences in Human Summary 1 = 17

Total Sentences in Human Summary 2 = 18

Sentences Common between System and Human Summary 1 = 7

Sentences common between System and Human Summary 2 = 9

Average Common Sentences = 8

(across all summaries) Average Human Summary Sentences = 17.5

(across all summaries)

Precision(P)=8/20=0.4

Recall(R) = 8/17.5=0.457

F-score = (2 .0*0.4 * 0.457)/(2 .0*0.4 *0.457)= 0.43

ROUGE-N RESULTS

The ROUGE-N scores obtained by the system over 3 different topics by calculating summaries from 10 articles each were as follows:

Rouge-N	Average R	Average P	Average F
1	0.37137	0.423	0.3934
2	0.17927	0.29161	0.22204

Table 3: Results across various n-gram values

DISCUSSIONS AND COMPARISONS WITH

STATE OF THE ART

COMPARISONS:

The result is compared with Rouge-1 and Rouge-2 results with competing systems.

Author	Average R	Average F	Technique used
[Lin and Hovy, 2000]	0.3935	0.3890	Latent Semantic Indexing
[Mihalcea and Tarau, 2004]	0.3733	0.3743	Graph based clustering
Best Possible	0.4003	0.4003	Gold standard summaries
Proposed System	0.37137	**0.3934**	DBSCAN with OPTICS Clustering

Table 4: Results comparison for ROUGE-1 values

Author	Average R	Technique used
Sripada et. al	0.0535	Sentence Ranking and TFIDF
[Amato et al., 2016]	0.15861	only OPTICS clustering
Proposed System	0.**17927**	DBSCAN with OPTICS Clustering

Table 5: Results comparison for ROUGE-2 values

DISCUSSIONS:

As shown, the system performs at-par if not better with state of the art competing systems In ROUGE-1 evaluations, it is equivalent to calculating total common words which exist in a sentence with respect to the corresponding sentence in reference summaries. In ROUGE-2, every 2-pair of words is taken and its existence is compared in the reference summaries. All the comparison Systems use extractive algorithms for generating summary and have been compared on the dataset in question.

The system shows high competency because:

- It does not need to use a sentence generator to generate summaries.
- The process of representing the sentence as a group of Triplets saves a lot of time and effort thus reducing computations required by algorithms which use sentence ranking and graphing algorithms.
- The system uses DBSCAN on the cluster ordering provided by OPTICS thus it becomes more accurate to validate the clusters formed.

CONSIDERATIONS:

- The competing systems use 10 topics for evaluation while this system used only 3 topics for evaluation due to restrictions on free availability of data.
- After preliminary analysis which involved visual inspection of summaries created by this system on queries generated by the user, the results will not waver from what is observed from 3 topics or 30 articles.
- The system is built such that it is equipped to handle even 100 articles and the summary size depends upon the number of articles it reads and is always coherent due to the chronological ordering of sentences before sentence selection.

V. CONCLUSION and Future Scope

This work presents a system to automatically collect, collate and summarize online newspaper articles based on a user submitted query. The dataset used for summarization is ad-hoc and is generated on-the-fly. Using pre-processing steps like Sentence tokenization, NER, Stemming and lemmatization and Triplet formation; the articles are broken down into manageable semantic atoms which are then clustered based on their semantic similarity. Results of the model were evaluated qualitatively on the DUC 2001 dataset The average F and average R score are 0.3934, 0.17927 respectively. These results show that implemented method efficiently generates extractive summaries with an Average F-score of 0.179 on ROUGE-2. The future work extended would be to include different media types like image, video audio etc., to try different and more advanced density based clustering techniques or deep learning neural networks to get better results, to employ various parallel programming techniques, multi-threaded approaches to increase the speed of summary generation.

References

[1]. McKnight, W. "Text data mining in business intelligence", Information Management, 15(1):80, 2005.

[2]. Barzilay, R. and McKeown, K. R., *"Sentence fusion for multidocument news summarization"*, Computational Linguistics, 31(3):297-328, 2005

[3]. Nenkova, A., Vanderwende, L., and McKeown, K., *"A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization"* In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pages 573-580. ACM, 2006

[4]. Lin, C.-Y. and Hovy, E., *"The automated acquisition of topic signatures for text summarization"*, In Proceedings of the 18th conference on Computational linguistics-Volume 1, pages 495-501. Association for Computational Linguistics, 2000

[5]. Bian, J., Yang, Y., and Chua, T.-S, *"Multimedia summarization for trending topics in microblogs"*, In Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, pages 1807-1812. ACM 2013.

[6]. Hennig, L. and Labor, D., *"Topic-based multidocument summarization with probabilistic latent semantic analysis"*, In RANLP, pages 144-149, 2009

[7]. Massandy, D. T. and Khodra, M. L., *"Guided summarization for Indonesian news articles"*, In Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of, pages 140-145. IEEE, 2014.

[8]. Mani, I. and Bloedorn, E., *"Multi-document summarization by graph search and matching"*, arXiv preprint cmp-lg/9712004, 1997.

[9]. Amato, F., d'Acierno, A., Colace, F., Moscato, V., Penta, A., and Picariello, A., *"Semantic summarization of news from heterogeneous sources"*, In International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pages 305-314. Springer, 2016.

[10]. Alruily, M., Ayesh, A., and Zedan, H., *"Crime profiling for the Arabic language using computational linguistic techniques"*, Information Processing & Management, 50(2):315-341, 2014.

[11]. Kiss, T. and Strunk, J., *"Unsupervised multilingual sentence boundary detection"*, Computational Linguistics, 32(4):485-525, 2006.

[12]. Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B., *"The penn treebank: annotating predicate argument structure"*, In Proceedings of the workshop on Human Language Technology, pages 114-119. Association for Computational Linguistics, 1994.

[13]. Tjong Kim Sang, E. F. and De Meulder, F., *"Introduction to the conll-2003 shared task: Language-independent named entity recognition"*, In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pages 142-147. Association for Computational Linguistics,2003

[14]. Nadeau, D. and Sekine, S, *"A survey of named entity recognition and classification"*, Lingvisticae Investigationes, 30(1):3-26, 2007.

[15]. Google (2017). Google Search Engine overview.