Digital Innovations
& Contemporary Research in SCIENCE,
ENGINEERING & TECHNOLOGY

Vol. 6. No. 1, March, 2018

Article Citation Format
O.P. Akomolafe & Mustapha Abdulqadir (2018). Multi-Document Text Summarization Using Sequential And Adjacency Information Model.
Journal of Digital Innovations & Contemp Res. In Sc., Eng & Tech.
Vol. 6, No. 1. Pp 121-134

Article Progress Time Stamps

Article Type: Research Article
Manuscript Received: iSTEAMS Extended Paper
Review Type: Blind
Final Acceptance:: 23rd March, 2018
DOI Prefix: 10.22624

Multi-Document Text Summarization Using Sequential And Adjacency Information Model.

O.*P.* Akomolafe *&* Mustapha Abdulqadir
Department of Computer Science
University of Ibadan
Ibadan, Nigeria
akomspatrick@yahoo.com, qadircmp@yahoo.com

ABSTRACT

Multi-document summarisation is the process of producing a single document from a collection of related documents, which help users to find required information quickly without wasting time in reading through set of documents. Existing work on multi-document text summarisation assumed that input documents all have the same text structure and format. This enables us to develop an expandable system that can be used to understand multiple document formats. The general architecture of the automatic text summarisation system is divided into two main modules, document selection and document summarisation. Document selection can be achieved using various tools such as indexing, tokenisation, and stemming and stop-words removal. The document summarisation module differs from one summariser to another based on the summarisation approach used. The summarisation module is made of five main processes, sentence splitter, sentence matcher, sentence selector, fusion and sentence ordering. The summary consist of several sentences from different documents, then the problem is how to arrange these sentences in order to produce a comprehensive content of the entire documents. In this paper, both adjacency and sequential information of the sentence were used. This approach was evaluated in term of similarity and ordering of the summary content. The results generated showed that there is higher similarity between manual by three different experts and system generated summary, also using Kendall's evaluation method showed a better output when compared feature-Adjacency model but equal Cluster-Adjacency model.

Keywords: Text Summarisation, Document formats, Sentence ordering Sequential Model, Cosine Similarity, Kendall's Model

1. INTRODUCTION

Text Summarization (TS) is a way of filtering essential gist of text in a document or set of related documents and conveying on it in concise form than the earliest form. On a Central level, TS is feasible due to the commonly happening repetition in content and in light of the fact that essential information is spread arbitrarily in text document. Sentence organization, identification of redundant sentence is a challenge that has not been completely determined at present. A summary can be utilized in an indicative way as a pointer to a few sections of the original text document, or in an informative approach to cover all applicable information of the text document. In both cases the more critical point of interest of utilizing summary is its lessened reading time[9]. Text Summarization strategies can be grouped into extractive and abstractive summarization.

An extractive summarization is system which includes choosing vital units, for example, sentences, sections and so on from the original document and summing them into condensed form. The importance of sentences is decided based on some statistical and linguistic feature of sentences. An Abstractive technique of summarization attempt to add to a comprehension of the key thoughts in a document and afterward pass on those ideas in straightforward language. It utilizes linguistic systems to analyse and interpret the content and afterward to discover the new ideas and representations to best depict it by producing another shorter document that passes on the most essential information from the original report [15].

With enormous documents formats providing same information (same topic), the available summarizers were unable to handle set of documents of related topic with different document formats (doc, docx, pdf, html, htm etc). To ensure an optimal problem free system, it is often a good idea to implement standard based technologies.

In Multi-document summarization position of sentence in the summarised document is significant, to provide a sensible meaning. Since a good summary should be fluent and readable to reader. Hence, ordering of sentence which organises text intofinal summary could not be ignored.
In this paper, a new concept is present for ordering sentences in the summary using adjacency - sequential information of the sentences and cross-format multi-document summarisation. It fits the following scenario, "User is faced with collection of different documents formats of related information on same subject. The system employs a means for making various document formats structure possible to avoid problems such as increase in error co efficient, waste of computational time to filter and rank sentences and synthetic filtering becomes impossible with increase in ambiguous information in the process". Secondly, like the feature-adjacency which focus main on adjacency and cluster-adjacency based maps group of sentence to a theme in the original documents by semi-supervised classification method and the adjacency pairs of sentences is learned from adjacency cluster the sentence belongs in summary.Itjoinboth adjacency and sequential information to ensure that the summary is more comprehensive.

2. RELATED WORKS

A variety of summarization method have been developed which share some basic features as describe as follow:

Ordering of Sentence appears to a large extent harder for multi-document summarization than for Single-document summarization [1][12]. The major basis is that compare with single document, multi-documents don't provide a natural arrangement of texts to be the basis of sentence ordering judgment. This is more obvious for sentence extraction based summarization systems. Multi-documents add sentences with both different authors and in writing styles, which means original documents could not directly offer orderingmeasure in multi-document summarization task.

Below are available methods of ordering. An approach of machine learning to study the combination of chronological, probabilistic, topic relatedness which yielda good result than having individual method[2]. Nie et al [16] adjacency value between sentence pairs to order sentences which is calculated based on adjacency features pair within the sentence pairs. Adjacencies between two sentences denote how closely they should be together. JiD. and Nie Y.[7] proposed clustered based method to sentence ordering for multi-documents summarization where each sentence are mapped to a topic in a source documents by a semi-supervised method and adjacency of pairs of sentences learned from the original documents based on the adjacency of group those sentence fit in. Lapata M. [12] offers Conditional probability of sentence pairs was applied to organize sentences. The conditional probabilities of sentence pairs were learned from a training corpus. With conditional probability of each sentence pairs, the approximate optimal global ordering was achieved with a simple greedy algorithm. The conditional probability of a pair of sentences was calculated by conditional probability of feature pairs occurring in the two sentences. The experiment results show that it gets significant improvement compared with randomly sentence ranking.Xiaoyan [19] approach integrates ranking and clustering by equally and concurrently updating each other so that the performance of both can be enhanced. Although, existing cluster based ranking approaches applied clustering and ranking in isolation.

Gongfu P. et al. [5] presented a realistic method of sentence ordering in extractive multi-document summarization tasks of Chinese language, by using Support Vector Machine (SVM) and classify the sentences of a summary into several groups in rough location according to the original documents. They adjust the sentence sequence of each group according to the opinion of directional relativity of adjacent sentences, discover the sequence of each group and later connect the sequences of different groups to generate the final order of the summary.

Yang-Wendy Wang [20] proposed and implements a procedure that combines constraints from query order and topical relatedness in human produced summaries of multiple documents in response to multiple questions. He tested the effectiveness of the constraints and construct a novel query-based quantity from the human produced summaries for the Document Understanding Conference (DUC) 2006 evaluation, after which he conducted an experiment using an automatic evaluation method based on Kendall's to evaluate and compare the electiveness of the method used where he concluded that better in ordering performance was achieved. The system only run the two constrains independently to determine ordering of sentence which reduces the system performance.
Chinese Automatic Text Summarization system for mobile devices an approach which uses both Automatic News Collection and Automatic Text to analyse target websites and find out the rule of the articles text in HTML page for Summarization [13].

Dragomir R [4] presented a web based multi documents and recommendation system called WebInEssence which was designed to provide users every opportunity to reduce the information overload and do simplified yet effective search and navigation. The system is possesses feature such as Ease to customize and personalize, High quality search, Scalable to handle multiple users. The system is flexible and can perform operation such as generic search, summarisation, clustering and personal mode, but fail to addressing reordering of the generated summary.

Table 2.1: Document Formats handled.

Extension	Name of file format
.doc	Word 97-2003 Document
.docx	Word Document
.docx	Strict Open XML Document
.htm, .html	Web Page, Filtered
.pdf	PDF
.txt	Plain Text
.xml	Word 2003 XML Document
.ppt	Power point Document
.pptx	Power point Document
.xlsx	Strict Open XML Spreadsheet
.xls	Excel 97-Excel 2003 Workbook

3. METHODOLOGY

The below architecture shows the text extraction process from the inception of uploading the documents. Basically, the document format have to be identify (format recognition) follow by unzipping the document for text extraction before the actually summary begins.

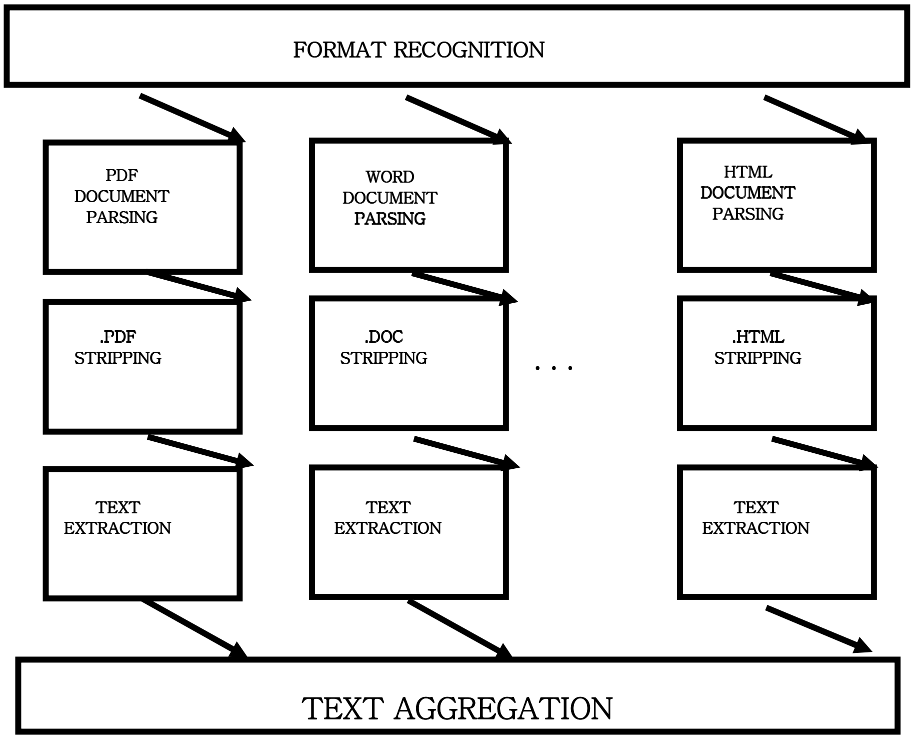

Figure 3.1: Format Identification and Text Extraction

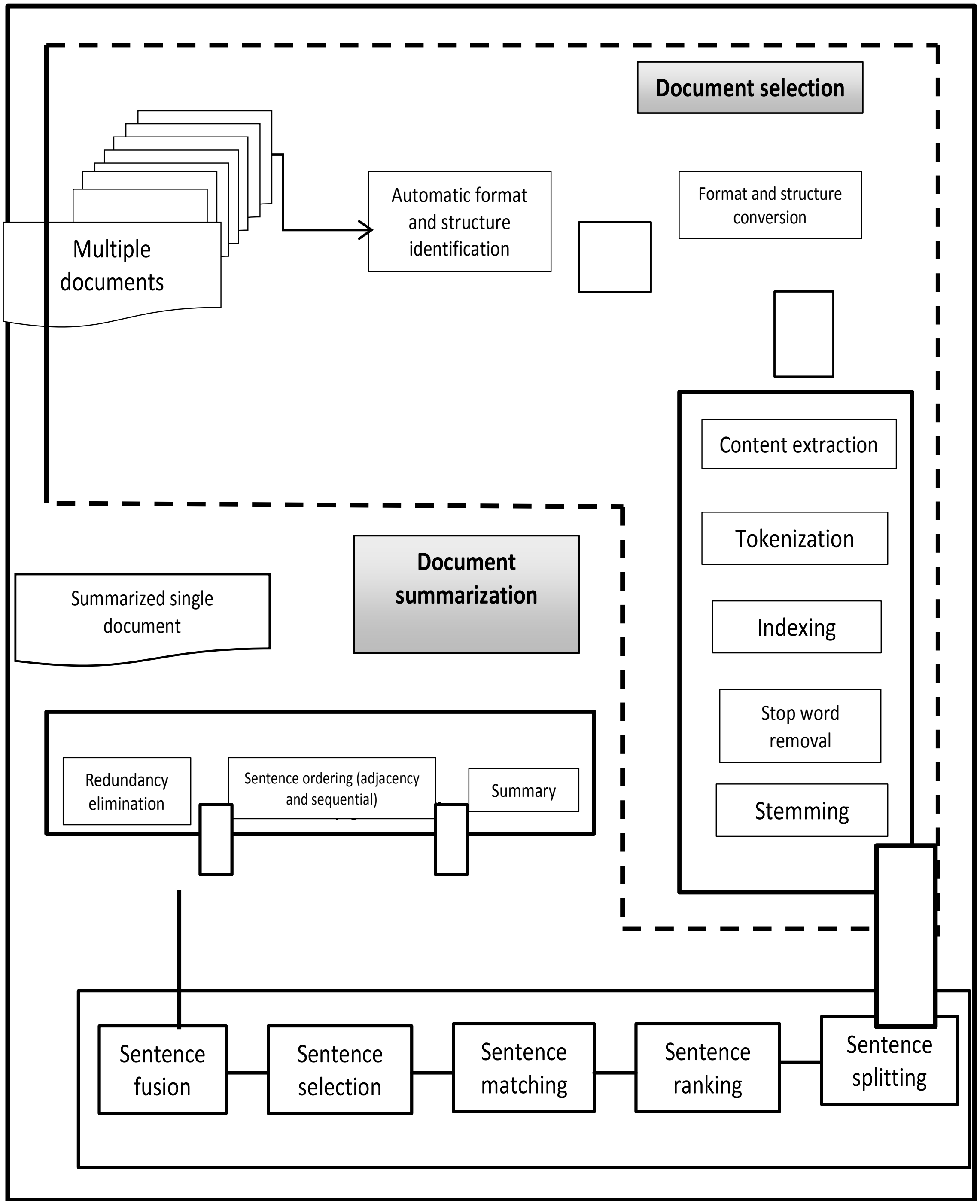

Figure 3.2 General Architecture of Multi-Documents Summarization

3.1 Document Summarization

The general architecture of the automatic text summarisation system is divided into two main modules, document selection and conversion and document summarisation. Document selection and conversion can be achieved using various tools such as document retrieval system. The document summarization module differs from one summariser to another based on the summarisation approach used. The summarisation module is made of five main processes, sentence splitter, sentence matcher, sentence selector, fusion and ordering. The five processes together are responsible in generating a single or multi-document summary. Prior to the summarisation process the data collection used needs to undergo processing steps that include conversion, indexing and applying natural language processing tools, such as indexing, tokenisation, stemming and stop-words removal. This will help in finding relationships between words and sentences in order to help ranking sentences and reducing redundancy in the case of multi-document summarisers. All the previous steps and processes are crucial for generating single and multi-document summaries.

i. **Sentence Matching:** This involves match a document's sentences to a certain information extracted from the document itself (highly ranked sentence or document's title), this could be the document's title or the highest ranked sentence in the document.

ii. **Sentence ranking:** involves assigning weight to each words contained in the sentence. It reflects the relative importance of the extracted tokens in documents. The most common models to compute terms weight is the tf.idf weighting, which makes use of two factors: the term frequency tf and the inverse document frequency *idf* [18]. The weight w of term j in a document i can be computed as.

$$Wij = tfij * idfj \qquad \text{Where} \qquad idfj = \log (n / dfj) \qquad (1)$$

And *tfij*, the term frequency, is the number of times that term *j* occurs in document *i,n* is the number of documents in the collection, and *dfj* is the document frequency of term*j* [6].

iii. **Selecting Selection:** involves choosing the sentence that will actually make the summary after matching.

iv. **Sentence fusion:** Sentence fusion is the task of taking two sentences that contain some overlapping information, but that also have fragments that are different. The goal is to produce a sentence that conveys the information that is common between the two sentences, or a single sentence that contains all information in the two sentences, but without redundancy.

v. **Sentence Ordering:** an integrated strategy for arranging sentences in the generated summary based adjacency and sequential of sentences to create coherent information. This is achieved be placing the adjacent sentences (close in proximity) and consider them sequentially considering the position occupy by each sentences in their originals document.

The steps involve in sentence ordering are listed below:

```
For each summary sentence
sentence1
        Foreach summary sentence
        sentence2
        If sentence1== sentence2
        Continue;
        End if

calculateAdjacency (sentence1,sentence2)
        End foreach
        SumAdjacency(sentence1)
end   foreach
sort sentence based on adjacency value
sort adjacent sentence sequentially(position) value
```

The adjacency between two sentences as in calculate Adjacency function is shown in thepseudo code below*:*

```
Begin
Tokenize sentence1
Tokenize sentence2
Return number intersection between the two sentences token.
End.
```

4. EXPERIMENTS

The implementation builds on the architecture that was presented in the methodology and tests the results against other results available on the internet. The system was tested on total number of one hundred and fifty multi-lingual texts documents having an average of 250 words. The text documents are of fifteen (15) categories, where each category contains ten related documents. The categories are legend personalities, such as writer, patriot, singer and sports personalities, different technical reports, and different newspaper articles on sports, politics, different travel narrations and short stories.

The frequency of some terms in summaries are taken different to see the efficiency of system in different cases and compared with the expert's summaries. First, each category of text documents summarized by 3 different experts. At the same time, the texts are summarized by the system. Then the results are compared using the cosine similarity algorithm was used to compute the relevance of the documents to the highest ranking word (term).

4.1 Evaluation

We evaluate the work using similarity measures algorithm to compute the similarity between the system and 3 different experts' summaries.

- Log frequency weighing = 1 + log (TF) (2)

Cosine Similarity (System, Human) = Dot product (A,B) / ||A|| * ||B|| (3)

Where

Dot product (A,B) = A[0] * B[0] + A[1] * B[1] * ... * d1[n] * d2[n] (4)

$||A||$ = square root (A $[0]^2$ + A $[1]^2$ +.+ A $[n]^2$) (5)

$||B||$ = square root (B $[0]^2$ + B $[1]^2$ + ... + B2 $[n]^2$) (6)

Term frequency (TF): Measure of number of times a term (word) occurs in a document. This involves breaking the text (document content) into tokens, then the token having the highest frequency is consider important to be included in the summary. Table below shows terms and their frequency on each token present in collection of related document been summarized.

Table 4.1: Term frequency (TF)

S/No	Terms	System	Expert 1	Expert 2	Expert 3
1	Tsunami	6	4	1	2
2	Earthquake	3	3	3	2
3	Romanian	6	3	1	1
4	Million	2	1	2	1
5	Women	1	5	2	2

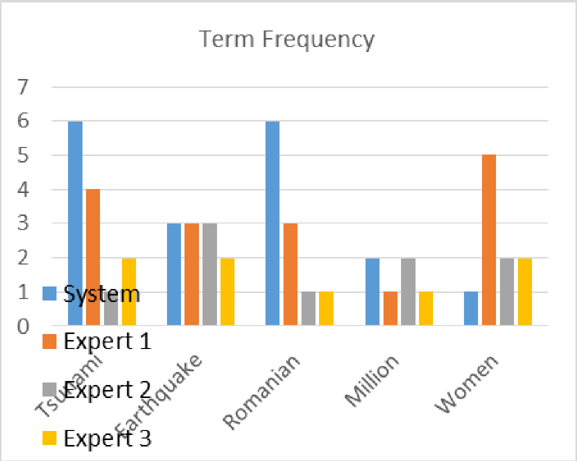

Figure 4.1: Term Frequency

Table 4.3 shows the number of occurrences of five terms (Tsunami, Earthquake, Romanian, Million, and Women) in each of the three Experts and the system summaries: Of course, there are many other terms occurring in each of these summaries. In this example we represent each of these summaries as a unit vector in five dimensions, corresponding to these five terms (only); raw term frequencies are used at this point, with no *idf* multiplier.

Table 4.2: Log frequency weighing
Log frequency weighing = 1 + log (TF)

S/No	Terms	System	Expert1	Expert2	Expert3
1	Tsunami	1.7782	1.6021	1.0000	1.3010
2	Earthquake	1.4771	1.4771	1.4771	1.3010
3	Romanian	1.7782	1.4771	1.0000	1.0000
4	Million	1.3010	1.0000	1.3010	1.0000
5	Women	1.0000	1.6990	1.3010	1.3010

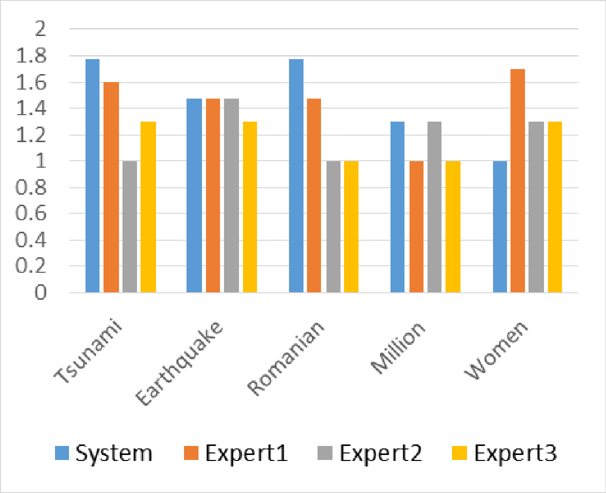

Figure 4.2: Log frequency weighing

Table 4.4: Log frequency weighing: the main purpose of doing a search is to find out relevant terms. In the first step all terms are considered equally important. In fact certain terms that occur too frequently have little power in defining the relevance. We need a way to weigh down the effects of too frequently occurring terms. Also the terms that occur less in the document can be more relevant. We need a way to weigh up the effects of less frequently occurring terms. Logarithms helps us to solve this problem.

Table 4.3: After length normalization

Normalization=B [0]/ square root (B $[0]^2$ + B $[1]^2$+ ... + B2 $[n]^2$) (7)

S/no	Terms	System	Expert1	Expert2	Expert3
1	Earthquake	0.5313	0.4784	0.3635	0.4890
2	Romanian	0.4414	0.4714	0.5369	0.4890
3	Million	0.5314	0.4714	0.3635	0.3759
4	Women	0.3888	0.2798	0.4729	0.3759
5	Tsunami	0.2988	0.4974	0.4729	0.4890

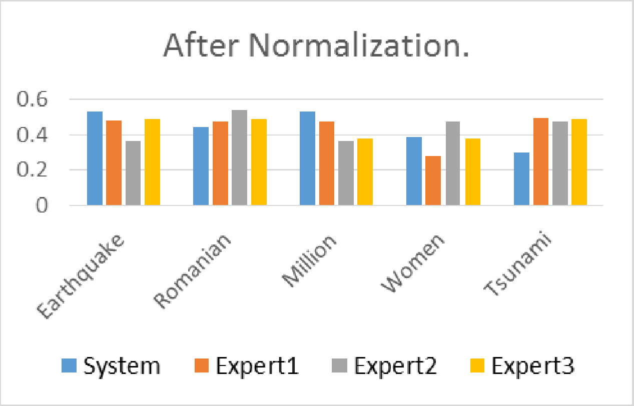

Figure 4.3.: After Normalization.

Cosine similarities between the system summary and experts summaries are computed as follows:

Cos (System, Expert1) = [(0.5313*0.4784) + (0.4414*0.4714) + (0. 5314*0.4714) + (0. 3888*0.2798) + (0.2988*0.4974)]

= [0.2542+ 0.2080 + 0.2505 + 0.1088+ 0.1486] = 0.9702

Cos (System, Expert2) = 0.9484.

Cos (System, Expert3) = 0.9677.

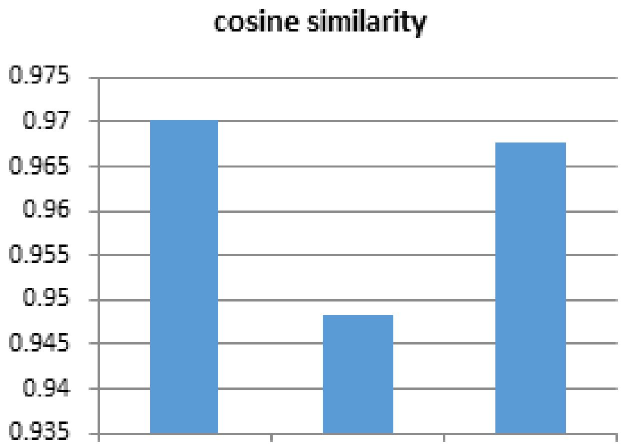

Figure 4.4: Cosine Similarities

Table 4.4: Number of words

	BEFORE	AFTER		
		Very Low	Low	Medium
NUMBER OF WORDS	2263	1207	804	441
% OF COMPRESSION	100	53.3	35.5	19.5

Table 4.5: sizes of summarized documents at different level

	BEFORE	AFTER		
		Very Low	Low	Medium
DOCUMENT SIZE (BYTE)	110877	14201	13086	11990
% OF COMPRESSION	100	12.8	11.8	10.8

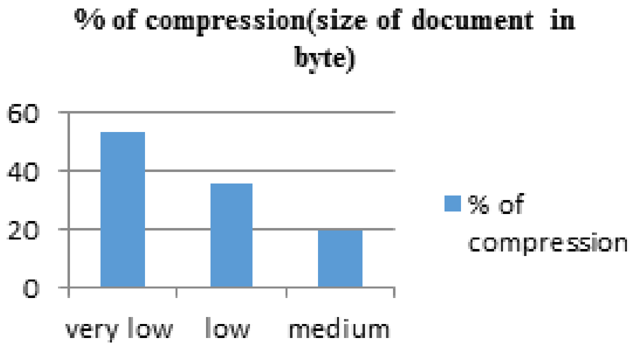

Figure 4.5: Compression rate

As seen in the chart, the accuracy of coherency of the system summary increases as percentage of both number of words and size of the summarized document increases.

$$Kendall^{'}s(\boxdot) = 1 - \frac{2(\quad)}{N(N-1)/2}\ldots \qquad (8)$$

Where N is the number of Sentences,NI is the number of interchange of adjacent sentence. Naturally, value ranges from -1 to 1, where 1 denotes that two ordering are same and -1 denotes completely converse ordering.Although, result is depend on the position occupied by a sentence in the original document.

Table 4.6: performance of the model using Kendall's

s/no	Natural ordering		□ Value
SA	A,B,C,D,E,F	A,B,C,E,D,F	0.73
FA	A,B,C,D,E,F	A,B,F,E,D,C	0.47
CA	A,B,C,D,E,F	B,C,A,D,E,F	0.73

5. CONCLUSION

This research presented two new concepts for multi-document summarization. First, it addresses the problem of cross-formats document (docx, html, pdf, etc) having related information for multi-documents summarization. Secondly, combine two sentence ordering methods such as the adjacency and sequential information between these sentences which yield a better result than individual methods. Finally, comparison between the system generated and manual summaries using cosine similarity measure. This measure ranges from 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity, to 1 meaning exactly the same, therefore from the graph above it depict that the system summary show high level of similarities with that of the experts summaries.

REFERENCES

[1] Barzilay R. McKeown K., Evans D., Hatzivassiloglou B., &Teufel, S. (2001). Columbia multi-doc Approach and evaluation. In Proceedings of DUC.

[2] Bollegala D., Okazaki N., Ishizuka I.” A machine learning approach to sentence ordering for multi-document summarization and its evaluation.” IJCNLP2005, LNAI 3651, pages 624-635, 2005.

[3] Dragomir R. R., T.Allison, et al “Evaluation challenges in large-scale document summarization”.

[4] Dragomir R. Radev, Weiguo Fan and Zhu Zhang “WebInEssence: A Personalized Web-Based Multi-Document Summarization and Recommendation System “School of Information Department of Electrical Engineering and Computer Science Business School University of Michigan Ann Arbor, MI 48103.

[5] Gongfu Peng, Yanxiang He, Ye Tian, Yingsheng Tian, and Weidong Wen “ANovel Method of Sentence Ordering Based on Support Vector Machine” Computer School,Wuhan University Wuhan 430079, P.R.China, pp.787- 794

[6] Grossman and O. Frieder. Information Retrieval: Algorithms and Heuristics. The Kluwer International Series of Information Retrieval. Springer, second edition, 2004

[7] Ji, Donghong and Yu Nie. 2004. Sentence Ordering based on Cluster Adjacency in Multi-Document Summarization. The Third International Joint Conference on Natural Language Processing, pp.745-750.

[8] Jimmy Lin,Dina Demner-Fushman“Evaluating Summaries and Answers: Two Sides of the Same Coin?”Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pages 41–48, Ann Arbor, June 2005.

[9] Joel Laroccca N., Alex A.F, Celso A.A.K “ Automatic text Summarization using Machine Learning Approach”, Pontifical Catholic University of Parama (PUCPR) Rua Imaculada Conceicao,1155.

[10] Josef Steinberger, Karel Jeˇzek (2009) “Evaluation measures for text Summarization”, Computing and Informatics, Vol. 28, 1001–1026, V (March).

[11] Karen Sparck J.”Automatic summarising: a review and discussion of the state of the art” January 2007.

[12] Lapata, Mirella. 2003. Probabilistic text structuring: Experiments with sentence ordering. Proceedings of the annual meeting of ACL, pp.545-552.

[13] Lei Yu, et al “A Chinese Automatic Text Summarization system for mobile devices” Faculty of Engineering, The University of Tokushima 2-1 Minamijosanjima, Tokushima 770-8506, Japan, 2003, pp. 426 – 429.

[14] McKeown K., Barzilay R. Evans D., Hatzivassiloglou V., Kan M., Schiffman B., &Teufel, S. (2001). Columbia multi-document summarization: Approach and evaluation. In Proceedings of DUC.

[15] Nabil Alami, Mohammed Meknassi, Noureddine Rais“ Automatic Texts Summarization: Current State Of The Art”, Laboratory of Computer and Modeling (LIM), University Sidi Mohamed Ben Abdellah (USMBA), Fez, Morocco, Journal of Asian Scientific Research, 2015, 5(1): 1-15

[16] Nie Yu, JiDonghong and Yang Lingpeng. An adjacency model for sentence ordering in multi-document Asian Information Retrieval Symposium (AIRS2006), Singapore. Oct. 2006.

[17] R. Barzilay, et al., "Information fusion in the context of multi-document summarization," in Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999, pp. 550- 557.

[18] Salton and M.McGill.Introduction to Modern Information Retrieval. McGraw-HillNY, USA,1986.ISBN0070544840

[19] Xiaoyan Cai Wenjie Li, You Ouyang, Hong Yan, "Simultaneous Ranking and Clustering of Sentences: A Reinforcement Approach to Multi-Document Summarization". Proceedings of the 23rd International Conference on Computational Linguistics, 2012, pp 134–142.

[20] Yang-Wendy Wang "Sentence ordering for multi-Document Summarization In response to multiple queries", Simon Fraser University Fall 2006,