9. EXPERIMENTATION & EVALUATION

To evaluate the model/framework developed, experiments are conducted to generate summaries of multiple documents automatically. The evaluation mechanism chosen and the corpus used are described in section 9.1 and the results of the summarization are reported in section 9.2.

9.1 Evaluation method adapted

One of the important decisions in evaluating text summarization is concerned with the choice of appropriate method and type of evaluation, since the goal of the present research is to automate the process of multi-document summarization that has been traditionally being done manually. The comparison of automatically generated extracts with those produced and rated ideal by humans, to provide a reasonable evaluation of the method. The comparison process needs to be automatic to eliminate any subjectivity. Thus the evaluation method takes the following form, the multiple documents for which the summary is to be generated is submitted to the system. The ideal summaries of the same set of documents are then presented along with the summary generated by the proposed system and the F-measure is computed. Since the main strength of the method lies in the ordering/sequencing of sentences in
the summary, ROUGE software have been used to automatically evaluate the generated summaries against the ideal ones.

9.1.1 Data Set

Most of the datasets/documents used for natural language understanding are domain specific such as newswire, medical, legal, financial etc. These documents have been used as standard reference for any NLP task. Therefore these standard data sets are used for experimentation in the proposed method.

NIST is an agency in the United States (US) commerce department, which publishes and annually conducts Data Understanding Conference (DUC) and workshops with themes on NLP tasks. The documents published in the year 2001 are newswire documents and have been published for a contest on event based text summarization. Therefore this dataset is chosen to not only evaluate the model but also compare the results with other related works.

DUC 2001 data sets contain a total of 308 documents. These documents are collected from English news agencies such as “wall street journal” and other English dairies. These sets have been further divided into 30 sets where each set representing a separate topic such as “The hurricane in Florida”. The number of documents in each of these sets ranged from 3 to 20. This dataset thus can support multi-document
summaries as well as single document summaries. The dataset also has a set of model summaries available along with these summaries. The model summaries are available for varying length namely 50, 100, 200 and 400. The model summaries are created by the human assessors of NIST. The contest was participated by event based summary model and hence keeping in mind all these characteristics. DUC 2001 data sets have been chosen for evaluation and corpus for the present work.

9.1.2 ROUGE Software

*Recall-Oriented-Understudy-for-Gisting-Evaluation (ROUGE)* is an automatic summary evaluation system which rates summaries on the characteristics of the cohesion and informativeness. It accepts the model summaries as inputs and then test summaries are compared to see how close the summaries are to the expert generated widely accepted model summaries. The score of the candidate summary is arrived in terms of F-measure as compared to the model summaries. The F-measure of the summary is defined as the cumulative similarity obtained by longest common subsequences (LCS) among individual sentences which are part of the model summaries and candidate summaries.

ROUGE (Lin chin Yew and Hovy 2003) is employed as a standard metric to compare the performance of the proposed model with that of other systems on same set of documents. The evaluation is done based on following scores
ROUGE-1...n: word n gram overlap between system summary with that of model summaries.

ROUGE-L: Longest common subsequence (LCS) (overlaps between human generated and machine generated summaries.

ROUGE W: weighted longest common subsequence (WLCS) which determines the length of consecutive matches encountered.

ROUGE-S: Skip bigram co-occurrence statistics measures skip bigrams between a reference and candidate translations.

ROUGE-SU: Same as ROUGE-S but by skipping uni-grams.

ROUGE is verified to be a good evaluation tool for multi-document summarization with a focus on content overlap. ROUGE is not a comprehensive evaluation method and intends to provide a rough description about the performance of machine generated summaries. It is found that ROUGE-2, ROUGE-L, ROUGE-W and ROUGE SU4 worked well for single document summarization while ROUGE-1 worked reasonably well for multi document summarization methods (Chin yew Lin 2004).

9.2 Results & Discussions

The quality of the summaries generated by our model is evaluated using ROUGE software. The F-measure scores for the proposed method lies from 0.25 to 0.56. In order to compare the proposed work with other
event based methods, analysis is made to see how well the performance of proposed model is better when compared with that of efficient event based method available. Lengths of the summaries varying between 50 to 400 are compared with other relevant event based extractive summarization methods [Filatova E et al 2003; Mingli Wu 2006] that have rated highest F-scores when compared with tf*idf, Adaptive greedy and static greedy methods. The comparative studies of these scores are tabulated in Table 9.1 and the results are found quite encouraging.

<table>
<thead>
<tr>
<th>% Time</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed Method better</td>
<td>58.6%</td>
<td>67.3%</td>
<td>76.5%</td>
<td>80.0%</td>
</tr>
<tr>
<td>other better</td>
<td>21.2%</td>
<td>15.2%</td>
<td>13.3%</td>
<td>13.4%</td>
</tr>
<tr>
<td>equal</td>
<td>20.2%</td>
<td>18.6%</td>
<td>10.2%</td>
<td>6.6%</td>
</tr>
</tbody>
</table>

*Table 9.1 Comparison of Proposed Method vs other methods*

**Length 400** word summaries generated by proposed method are analyzed and compared with same word length summaries of other models. It has been observed that when compared to other event based models out of 30 sets of documents, the quality of 24 sets of documents of proposed model was better than that of best known models [Filatova e et al 2003]. In terms of percentages it means that 80% of the times, the summaries generated by the proposed approach have improved quality. Out of the remaining six sets of summaries generated two sets are found to have equal quality with that of earlier known methods.

**Length 200** word summaries of proposed model are analyzed and it is shown that the summaries generated for 23 document sets out of 30
received better performances which resulted 76.5% on calculation of percentage. Out of the remaining summaries three sets are found to have equal quality with 10.2%.

For summaries of **Length 100**, it is observed that out of 30 sample scores taken individually for each set of document 21 sets are found to have better quality of summaries with proposed method and in terms of percentage it is denoted as 67.3%. Of the remaining, five sets are referred to have equal quality.

For summaries of **Length 50**, 17 to 18 sets of proposed method have ranked higher scores with a percentage of 58.6% times better than other event based methods while six document sets are found to have summaries of equal quality.

The experiments of table 9.1 showed that the performance of proposed approach is better for the summaries of longer length. This is true with any event based extractive summarization technique. The performance of the proposed method shown improvement for 80% of the cases is quite encouraging.

The **number of events selected is directly related to the length of summary**. As the events are more there is possibility of having more significant information in the summary generated and hence the performance of proposed system is shown to be higher for a summary generated with 400 words in length. Short summaries are abstractive where summaries are generated by small set of events. So in event based
approach the performance of ROUGE scores are higher for the summaries generated of longer length. Several experiments are performed by generating summaries of various lengths 50 words, 100 words, 200 words and 400 words. Evaluating these summaries by ROUGE software resulted in an average score value of 0.41. A sample calculation of ROUGE scores for the DUC data set is shown in figure 9.1.

![Figure 9.1: ROUGE Scores](image)

The graphs are plotted for ROUGE scores calculated for each document set where ROUGE scores form the Y-axis and number of document set forms the X-axis. These ROUGE scores are obtained for
different word length summaries, by having a comparison between “event-ordering using binary TCN” with that of best known related model.

The graph plotted in figure-9.2, depicts the ROUGE scores for the summaries generated by our Event based approach using TCN method compared with that of related work [Mingli Wu 2006]. These scores are generated when the length of the summary is 50. It is observed that the data sets numbered 11, 17 and 25 are scored higher for proposed method where ROUGE scores are approximately 0.4 and above. The document sets numbered <3, 5, 7, 9, 14, 21> are shown to have approximately equal ROUGE scores when compared with proposed method. The average F-score of 50 word length summaries is observed to be 0.29.

![Scores for Event based approach using Binary TCN Vs Other Methods](image)

*Figure-9.2 Comparison of ROUGE Scores for 50 word summaries*
The graph plotted in figure 9.3 depicted the comparison of ROUGE scores when the length of the summary is 100. It has been observed that the score of the document set 17 have increased from 0.42 to 0.54. The documents sets numbered 1, 2, 7, 14, 21 and 22 are found to have improved ROUGE scores for summary length 100 than when compared with 50 length summaries. The proposed approach are shown to have equal ROUGE scores for document sets 3, 5, 9, 26. Most of the scores are observed to have improved results with the increase in summary length from 50 to 100. The average F-score of 100 word length summaries is observed to be 0.39.

![ROUGE Scores for Event based approach using Binary TCN Vs Other Methods](image)

*Figure-9.3 Comparison of ROUGE Scores for 100 word summaries*

The graph plotted in the figure 9.4 gives the comparison of ROUGE scores when the length of the summary is 200. The results have shown that document set 7 have improved the score from 0.30 to 0.44 while
document set 9 have improved from 0.25 to 0.35. The ROUGE scores of the proposed method for documents sets numbered 4, 11, 16, 17, 19, 20 and 25 are found to have scores ranging above 0.45. On observation it is found that majority of the document sets have better scores on increase in the summary length from 100 to 200. The average F-score of 200 word length summaries is observed to be 0.44.

![Figure 9.4 Comparison of ROUGE Scores for 200 word summaries](image)

The graph plotted in the figure 9.5 gives the comparison of ROUGE scores when the length of the summary is 400. It can be observed that the proposed approach achieved significant improvement for ROUGE scores, on most document sets and had a highest score of 0.56 for document set 11. The scores of the proposed method for document set 19 have increased from 0.46 to 0.52 when the length of the summary is 400 words. It is clear that as the length the summary is being increased
reciprocally increased the scores of our method. The average F-score of 200 word length summaries is observed to be 0.52.

Figure 9.5 Comparison of ROUGE Scores for 400 word summaries

9.4 Summary

Event ordering in multi-document summarization technique using binary temporal constraint network has been experimented on DUC 2001 data set. The obtained summaries are evaluated using standard ROUGE evaluation software. The average F-measure obtained for proposed method is 0.41 which is shown to be higher when compared with related methods. Event ordering has helped in generating coherent summaries. The scores are increased reciprocally with increase in summary length. This is due to the fact that the proposed model belongs to the category of extractive event based summarization technique.