A Hybrid Approach for Extractive Multi-Document Summarization

1stNadeemAkhtar
DepartmentofComputerEngineeringAligarhMuslimUniversity
AligarhIndianadeemakhtar@zhcet.ac.in

2ndMMSufyanBeg
DepartmentofComputerEngineeringAligarhMuslimUniversity
AligarhIndiammsbeg@hotmail.com

3rdHiraJaved
DepartmentofComputerEngineeringAligarhMuslimUniversity
AligarhIndiahira.javed@zhcet.ac.in

Abstract—In this paper, we present a method for extractive multi-document summarization using a hybrid approach for sentence scoring that combines the benefits of regression model and topic model. Fusing the regression model based score with topic model based score combines the benefit of both methods for ranking sentences and words that are scored on the basis of both surface and topical features. We use support vector regression based model for obtaining sentence scores and sparseTLM topic model for obtaining word scores. Both sentence and word scores are combined using BiRank algorithm for sentence ranking. An Integer Linear Programming method is used to select summary sentences maximizing coverage of summary based on ROUGE scores. The proposed method is shown to outperform existing state-of-the-art methods.

Index Terms—Text Summarization, Topic Model, Support Vector Regression, BiRank Algorithm

1. INTRODUCTION

Extractive multi-document summarization systems often employ two-step process for selecting best summary for the document collection related to a single topic. First step is sentence ranking in which sentences of the document collection are ranked according to their importance or salience. In the second step, namely sentence selection, sentences are selected discarding the redundant sentences to maximize the final score of summary.

In the first step, importance score of sentences are predicted through either directly using some regression model [1], [2] or firstly scoring components (words, bigrams, phrases, entities) [3], [4] of sentences and then summing these component scores to rank sentences. Traditional regression models for both word, sentence scores use some query dependent, and some query independent handcrafted features to train a regression model. The features are mostly surface features like location, length, frequency etc. and does not include semantic information. Recently, deep learning based summarization methods use deep neural network to encode sentences into vector representations using semantic information [5], [6]. These neural network based methods show great improvement over traditional regression models.

Probabilistic topic models can explore thematic structure of large collection of documents by generating underlying topics, which are defined as probability distributions over words [7]. They provide means for extracting underlying themes in the text documents. Several topic models have been used for generating topic aware summary in MDS [8]. Topic models improve the extraction of general summary words for text summarization [9].

We propose to use a hybrid approach that integrates regression based and topic model based ranking approaches to combine the benefits of both approaches. In our proposed approach, we use a regression model for sentence score prediction and topic model for word score prediction. Regression based models have the benefit of predicting sentence importance by considering various surface features. For example, sentence position is a good indicator for sentence importance for summarization as first few sentences describe the main content of the document. Overlap with document title, number of named entities and stop word ratio in the sentence are also good features for sentence importance. These features are individual sentence features and does not consider the sentence relationships. The regression models cannot process the relationships among words that are necessary for the topical coverage of document collection. It may happen that top sentences found by the regression model belongs to a few topics leading to the poor topical coverage of the document collection. In extractive multi-document summarization, the sentence selection step either consider only top (50 to 100) sentences in optimization based methods like ILP formulations [10] or influenced by only the top few sentences in greedy methods like MMR [11] or submodular function optimization [12] based methods. As discussed earlier, the goal of the text summarization systems is to maximize coverage by including as many concepts (unigram, bigram etc.) as possible. Poor topical coverage of document collection leads to poor concept recall of the summarization systems resulting in lower ROUGE recall values.

The contribution of this paper is as follows.

1. We present a hybrid approach to integrate both sentence and word scores using a bipartite graph-ranking algorithm for scoring sentences and words.
2. We show that the obtained sentence rankings are better and produce better results for text summarization.
3. We also present a word based redundancy reduction method that does not require any tuning parameter and perform better than existing redundancy reduction methods.
4. We use our hybrid scores for query focused text summarization and show that obtained results are better than most of the state of the art methods.

2. PROPOSED APPROACH

We present a hybrid approach that fuses ranking scores obtained from a regression model and topic model. This combines the benefit of both models to rank sentences and words based on both surface and topical features. We use a support vector regression (SVR) based regression model for sentence ranking similar to the model used in [1]. We use sparseTLM for getting the word scores. To integrate both the regression model and topic model based scores, we use a mutual reinforcing ranking method BiRank [14] for bipartite graphs, which is similar to HITS [15], algorithm.

2.1 SVR model for sentence score

We use a support vector regression (SVR) based regression model for obtaining sentence scores similar to the model used in [10]. We use the same features as used in [10] with the addition of one new feature which measure the semantic similarity between the sentence and query using the concept of fuzzy bag of words [13], Sentence Query semantic Overlap (QSO). QSO is defined as follows:

$$QSO = \frac{\sum_t \sum_q Sim_{t,q}^{fbow}}{|Q|} \quad (1)$$

$$Sim_{t,q}^{fbow} = \begin{cases} COS(V[t],V[q]) & if\ COS(V[t],V[q]) > 0 \\ 0 & Otherwise \end{cases} \quad (2)$$

Where q and t are words of query and sentence respectively. Sim^{fbow} is fuzzy membership function defined on sentence and query words. V[t] and V[q] are word-embedding vectors for words t and q, which are obtained from word2vec model.

2.2 Topic Model for Word Score

We use sparseTLM [9] model to find the word score, which extends hPAM by incorporating spike and slab prior in the generative process of hPAM and pro-, vides improved high-level general topics resulting in better summary words for text summarization. We obtain both the unigram and bigram scores from sparseTLM as follows.

To obtain the word score, the probability of word given high-level topics in the generative process of document in sparseTLM is used. Assume p(w/h) is the probability of sampling word w given the high level topic h. H is the total number of high level topics. The word score S_w of word w is calculated as

$$S_w = \pi \sum_{h=1}^{H} \frac{p(\frac{w}{h})}{H} \quad (3)$$

Bigram scores are obtained from unigram scores. If both the unigrams in a bigram are nonstop-words, bigram score is set to the average of its two unigram scores. If anyone unigram is a stop-word, bigram score is set to the nonstop-word unigram score.

2.3 Integration of sentence and word scores using BiRank Algorithm

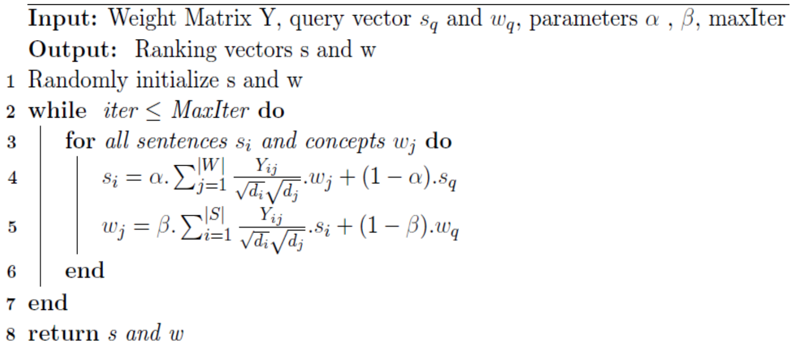
Since we have to integrate two types of scores i.e. sentence score and word score, we choose Birank [14], which is a bipartite graph-ranking algorithm. Birank can model the relationships between two types of entities based on their link structure of the associated bipartite graph as well as the prior or query information about the entities. Unlike the traditional random walk-based methods, BiRank optimizes a regularization function iteratively, converging the scores under the guidance of the query vector. The advantages of Birank for bipartite graph ranking are two-folds:

1. Unlike traditional random walk based ranking algorithm, BiRank smooths an edge weight symmetrically by the degree of its two connected vertices allowing edges connected to high degree nodes to be suppressed by normalization and reducing the effect of high degree nodes, which is the drawback of random walk, based diffusion networks [17]. This provide better quality results.
2. Birank also consider the prior beliefs or query about the nodes and incorporates them directly into the iterative ranking process.

Bipartite Graph Construction: To construct the bipartite graph, sentences and words of the document collection are considered as two types of nodes in the bipartite graph. An edge e_{sw} is drawn between sentence s and word w if sentence s contains word w. The weight Y_{esw} of edge e_{sw} is set to 1 as we con- sider all sentences and words equally initially.

Once the bipartite graph is constructed, final sentence and word scores are obtained using BiRank algorithm, which is depicted in Algorithm 1. In algorithm 1, s and w are sentence and word score vectors respectively. Vectors s_q and w_q are query vectors, which represents prior beliefs about sentence and word scores respectively. We initialize s and w randomly. Query sentence vector s_q is initialize to Scores i.e. the sentence score obtained in section 2.1 using our support vector regression model. Query word vector w_q is initialized to $Score_w$ i.e. the word score obtained in section 2.2 using sparseTLM topic model. The values d_i and d_j are degrees of sentence vertex s_i and word vertex w_j respectively connected through edge e_{siwj}. Parameters α and β controls the effect of initial query scores for sentences and words respectively. Value maxIter is the number of times BiRank iterations are repeated. Study shows that BiRank converges rapidly and requires 15 to 20 iterations for convergence. After execution of BiRank algorithm, final hybrid sentence and word scores are obtained in vectors s and w respectively.

In the above discussion, we integrate sentence and word (unigram) score into BiRank to get hybrid sentence and word (unigram) scores. In the same way, we can obtain hybrid bigram scores when we integrate bigram scores with sentence scores into BiRank algorithm.



```
Input: Weight Matrix Y, query vector $s_q$ and $w_q$, parameters $\alpha$ , $\beta$, maxIter
Output: Ranking vectors s and w
Randomly initialize s and w
while iter ≤ MaxIter do
    for all sentences $s_i$ and concepts $w_j$ do
        $s_i = \alpha.\sum_{j=1}^{|W|} \frac{Y_{ij}}{\sqrt{d_i}\sqrt{d_j}}.w_j + (1-\alpha).s_q$
        $w_j = \beta.\sum_{i=1}^{|S|} \frac{Y_{ij}}{\sqrt{d_i}\sqrt{d_j}}.s_i + (1-\beta).w_q$
    end
end
return s and w
```

Algorithm 1: BiRank Algorithm

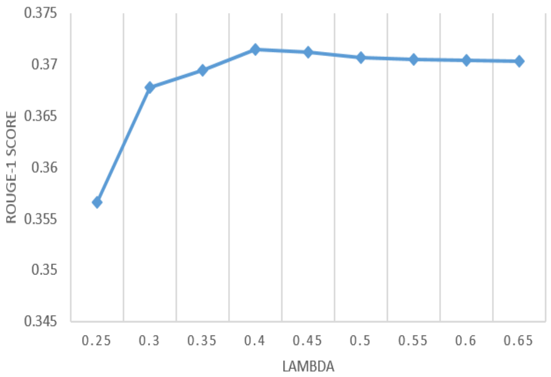


Figure 1: Variation of ROUGE-1 Score with λ for DUC2005 Dataset.

3. SENTENCE SELECTION APPROACHES

We use two existing summarization approaches MMR and ILP to show that hybrid scores obtained after integrating regression based and topic model scores improves the performance of these summarization approaches. We also propose a novel word-based maximal marginal relevance (MMR) method that does not require dependence on any threshold value unlike original MMR method.

3.1 Novel Word-based MMR Approach (MMR)

The framework of original MMR method for redundancy removal [2] is formulated as follows.

$$S_R(s|S) = \begin{cases} S_I(s) & \text{if } sim(s, S) \leq \lambda \\ 0 & \text{if } sim(s, S) > \lambda \end{cases} \quad (4)$$

S is the set of sentences in the current summary. $S_I(s)$ is the independent sentence score and sim(s,S) is the similarity between sentence s and current summary sentences set S which is usually defined as either maximum of cosine similarities between sentence s and any summary sentence s' in S or bigram overlap ratio. The drawback of MMR is that its performance heavily depends upon the value of threshold parameter λ. The variation of ROUGE-1 score with λ values for DUC2005 dataset is shown in figure 1. The value of λ is found by cross-validation on a validation set but it may not lead to best results. Ren et. al. [2] directly models the relative importance of a sentence s with respect to summary S in terms of relative importance scores of sentence s with respect to summary sentences s' in S instead of independent sentence score SI (s). Their framework as follows.

$$S_R(s|S)=min_s t\in S S_R(s|s') \quad (5)$$

S_R(s|s') is relative score of sentence s with respect to summary sentence s' . In this framework, there is no threshold parameter λ. For generating the summary, the sentence with best independent score SI (s) is selected as first summary sentence. Remaining sentences are selected iteratively using following greedy method.

$$S{*}=argmax_{s\in/S}min_{s}t\in S S_R(s|s')\qquad(6)$$

That sentence is added to the summary, which results in maximum relative increase in the summary score according to equation 6. Ren et. al. [2] shows that their method provides higher upper bound and better results for ROUGE scores without any need for tuning any threshold parameter.

Although the above framework provides better results than original MMR, there is a room for improvement. In equation 5, the relative score of sentence s with respect to entire current summary S is equal to the relative score of sentence s with respect to any summary sentence s' S_R(s|s') for which relative score of sentence s is minimum i.e. the relative score S_R(s|s') depends upon only one summary sentence instead of the entire current summary. It means that their solution is an approximate solution and has room for improvement. Peyrard [18] et. al. have specified a theoretical framework for finding the exact relative score and shown that it is an NP-hard problem. We observe that it is difficult to have exact solution in [18] because of the direct use of sentence scores. If we consider sentence scores based on word scores, then exact relative scores can be found easily.

Based on this observation, we present a word-based MMR method that finds the relative score of sentence s, which considers all the summary words. Our method use maximal Marginal Relevance over word scores instead of sentence scores. Sentence scores are obtained from words scores obtained in section 2.3.

$$S_I(s)=\sum_{w\in s}S(w)\qquad(7)$$

Firstly, the sentence having highest independent score SI (s) is included in the summary. All the remaining sentences s are assigned new relative scores as follows.

$$S_R(s|S)=\frac{S_I(s)-\sum_{w\in s\cap S}S(w)}{l(s)-|s\cap S|}\qquad(8)$$

Equation 8 shows that relative score of a sentence s, S_R(s|S), is sum of scores of only those words which are in sentence s and not in any summary sentence normalized by the effective length of sentence s. Effective length of sentence s is number of those words of s which are not in any summary sentence. The sentence having highest new relative score is added to the summary next. This procedure is repeated until desired summary length is achieved. From now onwards, we refer to our proposed approach as MMR, Ren's approach as MMRREN and traditional cosine similarity based approach as MMRCOS.

3.2 Existing Sentence Selection Approaches for Redundancy Reduction

Now, we describe existing sentence selection approaches that we used with our hybrid sentence and word scores for evaluation. We show in the section 5 that usage of our hybrid scores improved the performance of these existing summarization approaches.

MMRcos: In this approach, top ranked sentence is added to the summary first. Remaining sentences are re-ranked according to MMR criteria. If the cosine similarity of a sentence with any summary sentence is greater than a threshold value, its score is set to zero. The best of value threshold is found empirically.

Concept based ILP (ILP2): We use ILP formulation described in [19]. We maximize summary score, which is the sum of concepts (unigram/bigram) contained in the summary. The weights of concept are set as hybrid concept scores obtained in section 2.3. The objective function and associated constraints of the ILP formulation is as follows:

$$Max.Score_{sum}=\sum_{i=1}^{N}y_iw_i\qquad(9)$$

$$Subject\ to\colon\forall i\ s_i\,,w_i\in 0{,}1\qquad(10)$$

$$\sum_j s_j\,l_j\le L$$

$$\forall i,j\ \ s_j\,.Y_{ij}\ \le w_i$$

$$\forall i\ \ \sum_j s_j\,Y_{ij}\ge w_i$$

In the objective function equation 9, w_i is a binary variable corresponding to concept (unigram/bigram) i and y_i is its weight. s_j is binary variable corresponding to sentence j. l_j is length of sentence s_j. Y_{ij} indicates that concept i is in sentence j. Last two constraints ensure that if a concept is included then the sentence having it also included in the summary and if a sentence is included then all concepts contained in it are also included in the summary. We refer to this approach as ILP2.

4 EXPERIMENTS AND SETTINGS

In this section, we report the experiments performed and describe various experimental settings. We performed experiments to verify the following four propositions, which are about the efficacy of our proposed methods for extractive query focused text summarization.

Comparison of Sentence Rankings: The Integration of SVR sentence score in BiRank algorithm is more effective than using only SVR sentence scores for sentence ranking.

To verify this, we performed experiments using TopRank and MMRcos summarization method discussed in section 3.2. We chose these methods for this experiment because these two methods use sentence ranking scores for summary sentence selection.

We find two sentence rankings based on predicted ROUGE-

1 sentence scores using R1 regression model: first, using only SVR sentence scores and second, using hybrid word scores obtained from integration of SVR sentence scores into BiRank. R1 regression model is explained in section 4.2. For the second ranking, sentence scores are obtained from hybrid word scores using equation 7.

We evaluated both sentence rankings against the actual sentence ranking using overlap coefficient, Spearman coefficient and Kendall coefficient [20] intrinsically.

We considered only top 50 sentences to find overlap coefficient because performance of text summarization systems is affected by only top few sentences. For our ILP implementations, we also used only top 50 sentences. Overlap coefficient measures how many top 50 sentences are correctly extracted by the sentence rankings. The ranking which extract high ranked sentences or have high over- lap coefficient are likely to perform better for text summarization. Spearman coefficient is high if the two rankings are similar. Similarly, high Kendall coefficient means that two rankings are similar. For the evaluation of sentence rankings for text summarization, we reported ROUGE-1, ROUGE-2 and ROUGE-SU4 recall scores for the three DUC datasets.

Evaluation of Proposed Novel Word based MMR Method: The proposed word based MMR method for redundancy removal produces better results than existing MMR methods.

We compared our word based MMR method with Ren's MMR method MMR- REN and traditional MMR method MMRCOS based on cosine similarity. We implemented our MMR method with two types of word scores- first, when word scores are obtained using integration of SVR sentence scores into BiRank called MMRSVR and second, when both SVR sentence scores and Topic model word scores integrated into BiRank called MMRSVRTM. We also implemented traditional MMR method with two above mentioned scoring called MMRCOSSVR and MMRCOSSVRTM. We used R1 regression model for SVR sentence scores for this experiment. The ROUGE results for MMRCOS and its variants are found using best threshold value of λ. We reported ROUGE-1, ROUGE-2 and ROUGE-SU4 recall scores for query focused text summarization for DUC datasets.

Comparison of MMR and ILP2 with other Methods: Exploiting pro- posed hybrid scores for extractive multi-document summarization provides better ROUGE results than most of the current state of the art methods.

We chose our word based MMR and ILP2 methods for this experiments because they gave best performance. For MMR, we used both SVR sentence scores and unigram scores for integration into BiRank. R1 regression model is used for obtaining sentence scores. For ILP2, R1 regression model and unigram scores are used for obtaining ROUGE-1 recall and R2 regression model and bigram scores are used for obtaining ROUGE-2 recall scores as explained in section 4.2. We compared our word based MMR method and ILP2 method using our hybrid word scores with several state of the art methods. We selected best performing methods from each of the main text summarization categories for comparison. Following is the details of each method that we selected for comparison.

Now we discuss various experimental settings next.

4.1 Datasets

We used standard DUC datasets over query focused multi-document summarization task. Each document cluster has a query statement, which describes the details of the summary to be obtained. Each document cluster has four or nine reference summaries written by human experts. For support vector regression (SVR) training, ten largest (in terms of sentences) clusters of DUC2006 dataset are used and resulting model is used for sentence score prediction for DUC2005 and DUC2007 datasets. For DUC2006 dataset, ten largest clusters of DUC2007 dataset are used for SVR training.

4.2 SVR and BiRank Settings

For the implementation of SVR models, java library of LIBSVM is used. The SVR model parameters C and γ were set using grid based search.

We trained two models for independent sentence score prediction- one for predicting ROUGE-1 score and another for predicting ROUGE-2 and ROUGE-SU4 scores. For the first model, which we called R1 regression model, ROUGE-1 score of the sentence is used as target label. For the second model, called R2 regression model, ROUGE-2 score of the sentence is used as target label. The values of α and β in BiRank algorithm 1 were both set to 0.75.

4.3 ILP Environment Settings

All ILP formulations are written in GNU Mathematical Programming Language (GMPL) [21]. GMPL, a subset of AMPL, is an algebraic modeling language to describe high complexity problems for large scale mathematical computing. We used GLPK (GNU Linear Programming Kit) for solving ILP problem.

Table 1: Overlap, Spearman and Kendall Coefficients for Sentence Rankings

	DUC2005		DUC2006		DUC2007	
Coefficient	**Ranking1**	**Ranking2**	**Ranking1**	**Ranking2**	**Ranking1**	**Ranking2**
Overlap	0.3424	0.4772	0.3652	0.5336	0.3973	0.5667
Spearman	0.6287	0.7169	0.5355	0.7819	0.5312	0.7701
Kendall	0.4533	0.5238	0.3887	0.5783	0.3831	0.5629

4.4 Evaluation

We evaluate query focused summarization task on DUC2006 and DUC2007 datasets, which aim to generate 250 words summary. Output summary is evaluated using ROUGE [22]. We reported ROUGE-1 (recall against unigrams), ROUGE-2 (recall against bigrams) and ROUGE-SU4 (Recall with skip-bigram plus unigram with step 4) results with stop-words. For ROUGE score calculation, standard command in PERL ROUGE kit provided by the NIST is used.

5 RESULTS AND DISCUSSION

Now, we present and discuss the results of the each experiments.

5.1 Comparison of Sentence Rankings

The Integration of SVR sentence score in BiRank algorithm is more effective than using only SVR sentence scores for sentence ranking. The overlap, Spearman and Kendall coefficients obtained for the three DUC datasets are shown in table 1. Ranking1 is the sentence ranking obtained using only SVR sentence scores. Ranking2 is ranking obtained using hybrid word scores as described in section 4.

The results in table 1 show that ranking2 outperform ranking1 on all three correlation coefficients on all three DUC datasets. These results confirmed our belief that when SVR sentence scores are integrated into BiRank and hybrid word scores are used for sentence scoring, better sentence rankings are obtained. Higher overlap coefficient indicates that more top ranked sentences are extracted in top 50 sentences and higher Spearman and Kendall coefficients indicates that resulting sentence ranking is more similar to the actual sentence ranking.

5.2 Evaluation of Proposed Novel Word based MMR Method

The proposed word based MMR method for redundancy removal produces better results than existing MMR methods.

The results for the experiments on DUC2005 dataset is shown in figures 2. We observed that both variants MMRSVR and MMRSVRTM of our proposed method outperformed MMRREN and all variants of MMRCOS. The improvement for ROUGE-1 recall score for our MMR method were significantly better

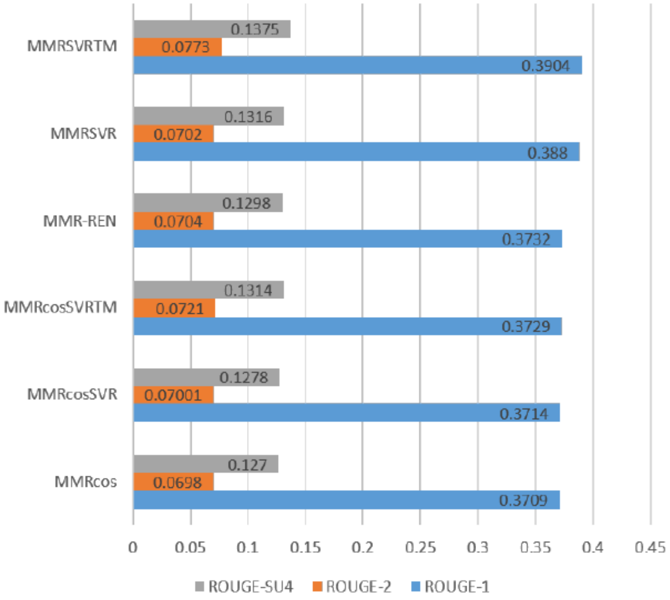


Figure 2: ROUGE Scores for various MMR Methods for DUC2005 Dataset.

than MMRREN. We observed a significant improvement of 0.0172 in ROUGE- 1 score for MMRSVRTM and 0.0148 for MMRSVR. We did not obtain any improvement in ROUGE-2 score for MMRSVR but obtain an improvement of 0.0071 for MMRSVRTM over MMRREN. The obtained results clearly strengthen our belief that our hybrid word score based MMR method performed more efficiently for redundancy removal in summary than Ren' MMR and traditional MMR methods.

We also observed that both MMRCOSSVR and MMRCOSSVRTM obtained better ROUGE results than MMRCOS, which also verify our belief that our hybrid word scores improves performance on text summarization.

Based on the results obtained, we list advantages of our proposed word based MMR method as follows.

1. Like Ren's method [2], our proposed method does not require any threshold value to be specified.
2. The relative score of a sentence depends upon the entire current summary providing exact relative score instead of an approximation.
3. The ROUGE results of our method for query-focused summarization are better than Ren's method MMRREN and traditional cosine similarity based MMR method.

5.3 Comparison of MMR and ILP2 with other Methods

Exploiting proposed hybrid scores for extractive multi-document summarization provides better ROUGE results than most of the current state of the art methods.

Table 2: Comparison of Summarization Methods for DUC2006 Dataset

Method	Rouge-1	Rouge-2	Rouge-SU4
MMR	0.4239	0.0985	0.1566
ILP2	0.4252	0.1046	0.1561
Peyrard[19]	0.4056	0.0926	0.1469
REN[2]	0.3984	0.0832	0.1414
Darakshaet.al[24]	-	0.08969	0.15070
SVR[1]	0.4018	0.0926	0.1485
Ranking-SVM[25]	0.4215	0.0983	0.1533
CES [26]	0.4300	0.0969	0.1563
MRC [27]	-	0.1094	0.1614
CTMSUM [28]	0.4157	0.0968	0.1548
HybHSum [29]	0.430	0.091	0.151
OCCAMS [30]	-	0.102	0.152
QODE [31]	0.4015	0.0928	0.1479
RSA-QFS [32]	0.4289	0.0873	-
ISOLATION [33]	0.4058	0.0896	-
AttSum [34]	0.4090	0.0940	-
VAEs-A [35]	0.3960	0.0890	-
SRSum [6]	0.4282	0.1046	-

The results for the experiments are listed in tables 2 and 3 for DUC2006 and DUC2007 datasets respectively.

Our methods MMR and ILP2 achieved good performance for three ROUGE metrics for all DUC datasets in general. The performance of ILP2 is better than MMR on all ROUGE metrics for the three datasets. MMR obtained significant improvements over most of the methods on ROUGE-1 metric but did not obtained improvement for ROUGE-2 and ROUGE-SU4. ILP2 performed better than most of the methods on all three ROUGE metrics.

ILP2 achieved better ROUGE scores than all regression and topic model based methods with one exception. Only HybHSum outperformed ILP2 on ROUGE-1 score for DUC2006 and DUC2007 datasets. For ROUGE-2 and ROUGE-SU4, ILP2 outperformed HybHSum. HybHSum also used both regression and topic model in its method but instead of considering word scores, it directly obtained sentence scores using regression performed on sentence features obtained using hierarchical LDA.

In comparison to optimization based methods, ILP2 outperformed all methods on ROUGE-2 and ROUGE-SU4 metrics for DUC2006 datasets but did not achieve better results for DUC2007. MMR is outperformed by all optimization based methods on ROUGE-2 and ROUGE-SU4 for all three datasets. We could not compare them on ROUGE-1 score as it is not provided by any optimization based method. Our both methods also performed better generally in comparison with most deep learning methods. SFSum is the best among all methods that we considered. It is benefited by its carefully designed deep network architecture. Other deep learning methods focused more on distributed representation of words rather than deep architecture. Our ILP2 outperformed all deep learning methods on all ROUGE metrics except SFSum, which is the best performer. RSA-QFS also outperformed ILP2 on ROUGE-1 for DUC2006 but outperformed by ILP2 for DUC2007. ILP2 achieved significant improvement over RSA-QFS on ROUGE-2 and ROUGE-SU4 for all DUC datasets. MMR achieved similar performance that it outperformed all deep learning methods except SFSum and RSA-QFS on ROUGE-1.

We also compared our results with Daraksha et. al. because they also used a bipartite ranking algorithm HITS on sentence-entity graph. Our ILP2 method achieved significantly better scores than Daraksha et. al. NCBSum-A is outperformed by ILP2 on all ROUGE metrics and by MMR by MMR on ROUGE-1. CES, a novel cross entropy method, achieved state of the art results on ROUGE-1. Our ILP2 is outperformed by CES on ROUGE-1 metric but achieved better scores on ROUGE-2 and ROUGE-SU4 for all datasets. MRC used a fuzzy hypergraph between sentence and topic nodes and achieved state of the art performance on ROUGE-2 and ROUGE-SU4 scores for DUC datasets for query focused text summarization. Our ILP2 method is outperformed by MRC on both ROUGE-2 and ROUGE-SU4 for DUC2006 and DUC2007 datasets. Overall, our ILP2 method performed consistently good for three DUC datasets. Its performance is better for DUC2006 than DUC2007 dataset. Among the state of the art methods that we used for comparison, ILP2 achieved second rank on ROUGE-2 for DUC2006 datasets and third rank on ROUGE-SU4 for DUC2006 respectively. It achieved fourth for ROUGE-1 for DUC2006. For DUC2007, it stood at third, seventh and sixth for ROUGE-1, ROUGE- 2 and ROUGE-SU4 respectively.

Table 3: Comparison of Summarization Methods for DUC2007 Dataset

Method	Rouge-1	Rouge-2	Rouge-SU4
MMR	0.4381	0.1100	0.1670
ILP2	0.4476	0.1223	0.1725
Peyrard[19]	0.4238	0.1087	0.1619
REN[2]	0.4255	0.1072	0.1621
Darakshaet.al[24]	-	0.1092	0.1673
SVR[1]	0.4342	0.1110	-
Ranking-SVM [25]	0.4461	0.1203	0.1701
PYTHY [36]	0.4260	0.1190	0.1680
Pingali et al. [37]	-	0.12448	0.17711
MRC [27]	-	0.1274	0.1792
NCBsum-A [38]	0.4289	0.1113	-
CES [26]	0.4543	0.1202	0.1750
hPAM [39]	0.4120	0.8900	0.1520
TTM [40]	0.4470	0.1070	0.1650
HybHSum [29]	0.4560	0.1140	0.1720
HierSum [41]	-	0.1180	0.1670
Galanis-ILP2 [10]	-	0.12517	0.17603
MCMR [42]	-	0.1221	0.1753
Lin et. al. [12]	-	0.12380	-
OCCAMS [30]	-	0.128	0.175
QODE [31]	0.4295	0.1163	0.1685
RSA-QFS [32]	0.4392	0.1013	-
ISOLATION [33]	0.4276	0.1079	-
AttSum [34]	0.4392	0.1155	-
VAEs-A [35]	0.4210	0.1100	-
SRSum [6]	0.4501	0.1280	-

Only SFSum outperformed ILP2 on all three ROUGE metrics on all three datasets. It is also outperformed by CES on ROUGE-1 and by MRC on ROUGE-2 and ROUGE-SU4.

6. CONCLUSION

In this paper, we have considered the task of query focused extractive text summarization. We have presented novel methods for both sentence scoring and sentence selection steps of query focused extractive text summarization. For sentence scoring, we have presented a hybrid approach to integrate sentence and word scores into a bipartite graph-ranking algorithm BiRank. We have shown that the hybrid scores improve performance of text summarization methods. For sentence selection, we have presented a word based redundancy reduction method that does not require any parameter tuning. We have established the following through our experiments.

1. Integration of SVR sentence scores into BiRank algorithm achieve better sentence ranking than using only SVR sentence scores.
2. Exploiting both SVR sentence and topic model word scores further improves scores and provide better ROUGE results for query focused text summarization.
3. The proposed novel word based redundancy reduction method obtain better ROUGE scores than existing redundancy reduction method for query focused text summarization.

4. The ILP2 method, which use the proposed hybrid scores, is among the best for query focused text summarization task for standard DUC datasets.

7. REFERENCES

[1] Y. Ouyang, W. Li, S. Li, and Q. Lu, “Applying regression models to query-focused multi-document summarization,” Information Processing & Man- agement, vol. 47, no. 2, pp. 227–237, 2011.

[2] P. Ren, F. Wei, C. Zhumin, M. Jun, and M. Zhou, “A redundancy-aware sentence regression framework for extractive summarization,” in Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 33–43, 2016.

[3] K. Hong and A. Nenkova, “Improving the estimation of word importance for news multi-document summarization,” in Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 712–721, 2014.

[4] M. Rei, “Semi-supervised multitask learning for sequence labeling,” arXiv preprint arXiv:1704.07156, 2017.

[5] R. Nallapati, B. Zhou, and dos Santos et. al., “Abstractive text summarization using sequence-to-sequence rnns and beyond,” in Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, pp. 280–290, 2016.

[6] P. Ren, Z. Chen, Z. Ren, F. Wei, L. Nie, J. Ma, and M. De Rijke, “Sentence relations for extractive summarization with deep neural networks,” ACM Transactions on Information Systems (TOIS), vol. 36, no. 4, pp. 1–32, 2018.

[7] D. M. Blei, “Probabilistic topic models,” Communications of the ACM, vol. 55, no. 4, pp. 77–84, 2012.

[8] R. Rani and D. Lobiyal, “An extractive text summarization approach using tagged-lda based topic modeling,” Multimedia tools and applications, vol. 80, no. 3, pp. 3275–3305, 2021.

[9] N. Akhtar, M. Sufyan Beg, and M. Muzakkir Hussain, “Sparse two level topic model for extraction of general summary words,” Journal of Interdisciplinary Mathematics, vol. 23, no. 1, pp. 303–310, 2020.

[10] D. Galanis, G. Lampouras, and I. Androutsopoulos, “Extractive multi- document summarization with integer linear programming and support vector regression,” in Proceedings of COLING 2012, pp. 911–926, 2012.

[11] J. G. Carbonell and J. Goldstein, “The use of mmr, diversity-based reranking for reordering documents and producing summaries.,” in SIGIR, vol. 98, pp. 335–336, 1998.

[12] H. Lin and J. Bilmes, “A class of submodular functions for document summarization,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 510–520, Association for Computational Linguistics, 2011.

[13] R. Zhao and K. Mao, “Fuzzy bag-of-words model for document representation,” IEEE Transactions on Fuzzy Systems, vol. 26, no. 2, pp. 794–804, 2017.

[14] X. He, M. Gao, M.-Y. Kan, and D. Wang, “Birank: Towards ranking on bipartite graphs,” IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 1, pp. 57–71, 2016.

[15] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” Journal of the ACM (JACM), vol. 46, no. 5, pp. 604–632, 1999.

[16] D. Parveen and M. Strube, “Multi-document summarization using bipartite graphs,” in Proceedings of TextGraphs-9: the workshop on Graph-based Methods for

Natural Language Processing, pp. 15–24, 2014.

[17] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly, “Video suggestion and discovery for youtube: taking random walks through the view graph,” in Proceedings of the 17th international conference on World Wide Web, pp. 895–904, 2008.

[18] M. Peyrard and J. Eckle-Kohler, “Optimizing an approximation of rouge-a problem-reduction approach to extractive multi-document summarization,” in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1825–1836, 2016.

[19] C. Li, X. Qian, and Y. Liu, “Using supervised bigram-based ilp for ex- tractive summarization,” in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 1004–1013, 2013.

[20] M. G. Kendall, “Rank correlation methods.,” 1948.

[21] A. Makhorin, “Modeling language gnu mathprog,” Relat´orio T´ecnico, Moscow Aviation Institute, vol. 63, 2000.

[22] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, pp. 74–81, 2004.

[23] D. Parveen and M. Strube, “Integrating importance, non-redundancy and coherence in graph-based extractive summarization,” in Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015.

[24] C. Shen and T. Li, “Learning to rank for query-focused multi-document summarization,” in 2011 IEEE 11th International Conference on Data Mining, pp. 626–634, IEEE, 2011.

[25] G. Feigenblat, H. Roitman, O. Boni, and D. Konopnicki, “Unsupervised query-focused multi-document summarization using the cross entropy method,” in Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 961–964, 2017.

[26] H. Van Lierde and T. W. Chow, “Learning with fuzzy hypergraphs: a topical approach to query-oriented text summarization,” Information Sciences, vol. 496, pp. 212–224, 2019.

[27] G. Yang et al., “A novel contextual topic model for multi-document summarization,” Expert Systems with Applications, vol. 42, no. 3, pp. 1340–1352, 2015.

[28] A. Celikyilmaz and D. Hakkani-Tur, “A hybrid hierarchical model for multi- document summarization,” in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 815–824, Association for Computational Linguistics, 2010.

[29] S. T. Davis, J. M. Conroy, and J. D. Schlesinger, “Occamsan optimal combinatorial covering algorithm for multi-document summarization,” in 2012 IEEE 12th International Conference on Data Mining Workshops, pp. 454– 463, IEEE, 2012.

[30] S.-h. Zhong, Y. Liu, B. Li, and J. Long, “Query-oriented unsupervised multi-document summarization via deep learning model,” Expert systems with applications, vol. 42, no. 21, pp. 8146–8155, 2015.

[31] T. Baumel, M. Eyal, and M. Elhadad, “Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models,” arXiv preprint arXiv:1801.07704, 2018.

[32] Z. Cao, F. Wei, L. Dong, S. Li, and M. Zhou, “Ranking with recursive neural networks and its application to multi-document summarization,” in Twenty-ninth AAAI conference on artificial intelligence, 2015.

[33] Z. Cao, W. Li, S. Li, F. Wei, and Y. Li, “Attsum: Joint learning of focusing and summarization with neural attention,” arXiv preprint arXiv:1604.00125, 2016.

[34] P. Li, Z. Wang, W. Lam, Z. Ren, and L. Bing, “Salience estimation via variational auto-encoders for multi-document summarization,” in Thirty- First AAAI Conference on Artificial Intelligence, 2017.

[35] K. Toutanova, C. Brockett, M. Gamon, J. Jagarlamudi, H. Suzuki, andL. Vanderwende, “The pythy summarization system: Microsoft research at duc 2007,” in Proc. of DUC, vol. 2007, 2007.

[36] R. K. Prasad Pingali and V. Varma, “Iiit hyderabad at duc 2007,” Proceedings of DUC 2007, 2007.