Extractive Summarization via Overlap-Based Optimized Picking

Gaokun Dai(✉) and Zhendong Niu

Beijing Institute of Technology, Beijing, China
dgkmao2340465@gmail.com, zniu@bit.edu.cn

Abstract. Optimization-based methods regard summarization as a combinatorial optimization problem and formulate it as weighted linear combination of criteria metrics. However due to inconsistent criteria metrics, it is hard to set proper weights. Subjectivity problem also arises since most of them summarize original texts. In this paper, we propose overlap based greedy picking (OGP) algorithm for citation-based extractive summarization. In the algorithm, overlap is defined as a sentence containing several topics. Since including overlaps into summaires indirectly impacts on salience, summary size and content redundancy, OGP effectively avoids the problem of inconsistent metric while dynamically involving criteria into optimization. Despite of greedy method, OGP proves above $(1 - 1/e)$ of optimal solution. Since citation context is composed of objective evaluations, OGP also solves subjectivity problem. Our experiment results show that OGP outperforms other baseline methods. And various criteria proves effectively involved under the control of single parameter β.

Keywords: Overlap-based optimization · Non-decreasing submodular objective function · Citation-based extractive summarization

1 Introduction

Moving to a new area is always painful for researchers, especially when the knowledge within the field becomes boosted and complicated. Researchers need to spend great amount of efforts on reading various papers for a deeper understanding. So the tool which summarizes papers with only several sentences could help researchers to study in a more efficient way.

To generate a good summary, various criteria should be taken into consideration. Typically summaries should contain the most salient contents while avoiding repetition or redundancy. Meanwhile, summaries should be concise and contain as many topics as possible. One of solutions is to employ centroid-based methods [8,9] where each topic represents a knowledge aspect of a paper. Then the summary is generated by selecting the most salient sentence per topic, which efficiently avoids redundancy while remaining salience. However, centroid-based methods would incredibly increase summary size if topics are numerous to cover.

A. Bouguettaya et al. (Eds.): WISE 2017, Part I, LNCS 10569, pp. 135–149, 2017.
DOI: 10.1007/978-3-319-68783-4_10

To solve this problem, a simple but effective solution is to select top-N sentences, sorted according to salience [29]. In this way, summary size is reduced while remaining content salience as much as possible and still guaranteeing low redundancy. Nevertheless, top-N strategy would ignore less salient topics, which leads to information loss. Since full topic coverage becomes urgent especially when the space of output is limited [20], top-N strategy is not always preferred.

Optimization-based methods try to involve all criteria into optimization process and pursue balance instead of inclination. Generally, the objective function is formulated as the weighted linear combination of criteria measurements. Then integer linear program(ILP)[27,28] is employed to get optimal solution. Nevertheless since obtaining optimal solution of the optimization problem with cardinality constraint is NP-hard, fast and performance guaranteed greedy algorithms [15,16,38] is preferred. Obviously weight setting is extremely important when measuring performance of greedy methods. However since criteria metrics are usually inconsistent, it is hard to set proper weight for each criterion in the linear combination.

In this paper, we propose Overlap based Greedy Picking(OGP) algorithm which formulates optimization process based on trade-off between overlaps and non-overlaps for citation-based extractive summarization tasks. In this method, overlap is defined as a sentence containing several topics while non-overlap is a sentence containing single topic. And we claim that including overlaps into summaries have following three impacts in detail.

1. Under the constraint of full topic coverage, less sentences are required if more overlaps included.
2. Since reiterating same topics is disgusting, overlaps are preferred especially when they can avoid topic repetitive coverage, which implies reducing content redundancy.
3. More overlaps might cause less salient summary since overlap is not always salient enough to represent every contained topic due to limited sentence length.

Thus formulated as the trade-off process between overlaps and non-overlaps, OGP indirectly involve various criteria and effectively avoids the problem of inconsistent metrics. Despite of a greedy method, OGP proves performance guaranteed due to the design of objective function mentioned in Sect. 3.

Generally there are two main contributions. Firstly we propose an effective optimization methodology that formulates summarization based on the trade-off between overlaps and non-overlaps, which indirectly involves criteria and avoids the problem of inconsistent criteria metrics. Secondly we design an objective function, which guarantees the performance of greedy solution. Further via a single parameter β in the objective function, balance status of criteria can be easily controlled.

The rest of this paper is organized as follows. We introduce related work in Sect. 2. The formulation of the constrained optimization problems are discussed in Sect. 3. Section 4 shows the detail of OGP. In Sect. 5, details about datasets, baseline approaches and evaluation metrics are presented. Then experiments results and relative discussion are described in Sect. 6. Section 7 concludes the main work in this paper and suggestions on future work.

2 Related Work

2.1 Citation-Based Summarization

Citances refer the text spans around the citations and contain information about the important facts in the cited paper [24]. [12,30] extend the definition of citances to including non-explicit text spans which talk about the cited paper but not explicitly expressed. Thus citation context, which is composed of all citances, provides a rich context about the important knowledge aspects of the cited paper and is suitable to be used for summarization [24]. [21] produce the impact-based summary of a single research paper based on the language modeling method, where citances are extracted to describe the impact. [29] finish citation-based single document summarization on scientific articles by applying clustering firstly to find different contributions of a target paper and then utilizing LexRank [5] for sentence selection. [25] prove the possibility of citation-based survey, where citation categorizations are proposed to obtain a survey. Further [22] compare surveys originated from abstracts, full papers and citation context, whose results indicate that multi-document technical survey creation benefits much from citations. Besides of containing enough information for summaries, citation context also help to solve the subjectivity problem, since citances are composed of objective evaluations from other scholars. Thus in this paper, we propose summarization method purely based on citation context.

2.2 Optimization-Based Approaches

Various summary criteria should be considered to generate qualified summaries. Salience is one of the most popular metric, which is often measured by sentence-level features such as position, TFIDF. Relying on graph theory, graph model attracts attention, where Random walk [4,5] is applied to assign centrality score to measure salience. Besides other graph theory technologies such as minimum dominating set model [34] and graph cut [31] are later applied to score salience. Nevertheless, most of these centroid-based methods pay little attention on limiting redundancy. Consider a scenario where a text contains one central topic and several other related topics. In order to gain salience, sentences containing the same central topics in this case are preferred, which definitely cause redundancy. A methodology for pruning redundancy is via clustering, where each group of similar sentences represents a single topic [7,19,29,35]. Picking the most central sentence per cluster could effectively avoid redundancy while remaining salience. We regard this cluster-based methodology deserving attention for its simple and effective way of restricting redundancy, even though most hard clustering methods ignore the existence of overlap as [36] states.

Optimization-based methods regard summarization as combinational problem, where objective function is formed as a weighted linear combination of various criteria metrics. Maximum Marginal Relevance(MMR) [2] cast redundancy penalty on the centrality score to reduce redundancy. [18] describe MMR as a knapsack problem and propose to utilize an integer linear program(ILP) solver to find the optimal solution. Complying with knapsack constraint, maximum coverage model [6,37] is proposed where sentences are picked to maximize information units coverage. To tackle coherence problem, dependency-based discourse tree [10] and graph model [27,28] is also employed, where ILP is used to get optimal solution. Since it often takes much time to get optimal solution, several performance-guaranteed fast approaches are proposed. [26] add redundancy-penalty constraint over the objective function and find feasible approximate solution based on Lagrange heuristics. [14] formalize the problem as submodular function maximization, which is solved via a simple greedy algorithm near-optimally. [15] extend the cardinality constraint to a general budget constraint and [16] points out monotone nondecreasing submodular function is an ideal objective function for optimization-based summarization. Then [17] introduces steps to design a more complicated submodular objective function using submodular shells. Later [23,38] proves submodular objective function also works well in the context of discourse structure. However we find most of current optimization-based methods mentioned above pay little attention on involving summary size into optimization. And naive top-N strategy is usually employed to comply with length constraint, which is indeed a post-processing style. Thus while top-n strategy violates the principle of optimization, we believe methods which dynamically involves various criteria including summary size is better.

3 Problem Formulation

In order to take various criteria dynamically involved as mentioned above, we formulate the citation-based extractive summarization task as the constrained optimization problem. And the problem is formally shown in following Eq. (1).

$$\begin{aligned} &Maximize. \sum_{n}^{N} Rep_n = \sum_{n}^{N} \{ \frac{\sum_{c}^{C} \varPhi_n(c) S_{nc}}{\varPhi_n} + \beta [\sum_{c}^{C} \varPhi_n(c)]^2 \} \\ &ST. \sum_{n}^{N} \varPhi_n(c) = 1, \forall c \in C \end{aligned} \tag{1}$$

Equation (1) shows the objective function of the optimization problem. n denotes a candidate sentences and N is the set of all sentences. c denotes a topic of articles and C represent all topics, which are explored by the method in Sect. 4.1. $\varPhi_n(c)$ equals 1 when candidate sentence n is selected and covers c, otherwise 0. $\varPhi_n$ refer the amount of topics that sentence n contains. Rep_n is the metric to measure sentence n, which is introduced further in Sect. 4.2. The first part of Rep_n is to measure the salience density where S_{nc} is the metric to measure salience of sentence n to cover topic c. The second part measures the amount of topics that sentence n covers. Note that under the constraint in Eq. (1), all topics should be covered, which guarantees full topic coverage. Thus the Rep_n of sentence n could be affected when some topics which n could cover are covered by others, because each topic should be covered by a single candidate sentence according to the constraint.

According to Eq. (1), the objective function is designed to maximize sum of *Rep* scores for a summary. And *Rep* reflects the performance of salience density and amount of covering topics while β is the relative importance of these two metrics. Generally compared to non-overlaps, overlaps contains more topics and have relatively lower salience density, as mentioned above. Thus when *Rep* focuses more on salience density, non-overlaps are preferred. Conversely when *Rep* focuses more on the amount of contained topics, overlaps are preferred. In fact, we claim that content redundancy is also implicitly involved in *Rep*. And the level of content redundancy is lower when overlaps are preferred, which is explained in Sect. 4.2. Thus based on the trade-off process between overlaps and non-overlaps under the control of β, Eq. (1) transforms extractive summarization task into the trade-off between criteria of salience, summary size and content redundancy.

4 Proposed Method

To solve the constrained optimization problem mentioned in Sect. 3, Overlap-based Greedy Pick(OGP) is proposed. Generally OGP could be defined as a three-step process. Firstly, a fast single-linkage hierarchical method [1] is employed to explore topics of the citation context. Sentences containing more than one topic are defined as overlaps. Secondly representativeness metric is proposed to enable the trade-off process between overlaps and non-overlaps. Finally, a performance guaranteed greedy algorithm is designed to solve the optimization problem and generate summaries.

4.1 Overlap Discovery

Topic exploration is the main task for overlap discovery, where topics are defined as sentences containing more than one topic. Then cluster-based method is employed to explore topics, where each cluster of sentences is defined as one topic. Since each candidate sentence could contain several paper topics, traditional cluster-based method which enforces one cluster per sentence is not suitable. Further, efficiency of the clustering method is very important especially when there is still much subsequent work to do. Then a fast single-linkage hierarchical clustering method [1] is employed in this paper.

[1] is a graph-based clustering method. So in order to apply this method, graph model is built firstly where each node represents candidate sentence and the edge denotes the corresponding sentence similarity based on typical TF.IDF metric. [1] treats links or edges as the cluster target instead of nodes and pursues

maximal links within each cluster. Then the overlap is the node with links belonging to different clusters, which indirectly solve the problem of being enforced to belong to only one cluster. Specifically, Jaccard Index [11] is employed to measure the similarity of link pairs as Eq. (2) shows, where e_{ik}, e_{jk} represent the edge between k,i and k,j respectively. Neighbors of node i are indicated by ni.

$$S(e_{ik}, e_{jk}) = \frac{|n(i) \bigcap n(j)|}{|n(i) \bigcup n(j)|} \tag{2}$$

Then single-linkage hierarchical clustering process is applied. Initially, individual cluster is assigned to each link. Cluster pairs containing link pairs with currently largest similarity, merged together at each step until all separate clusters merged into the single one. Partition density in Eq. (4) is recorded at each step to measure current clustering result. And finally the result turn to be the one with the highest partition density. Obviously this hierarchical merging mechanism indeed consumes very little time, which would not be increased sharply with the boosted data scale.

$$D_c = \frac{m_c - (n_c - 1)}{n_c(n_c - 1)/2 - (n_c - 1)} \tag{3}$$

$$D = \frac{2}{M} \sum_c m_c \frac{m_c - (n_c - 1)}{(n_c - 2)(n_c - 1)} \tag{4}$$

In Eq. (3) D_c indicates the link density of cluster c. The overall partition density D in Eq. (4) is then the average of D_c, weighted by the fraction of present links. m_c, n_c represent the number of links and number of nodes in this cluster. Thus D_c is in essence the number of links in cluster c normalized by the minimum and maximum numbers of links between those connected nodes. For example, cluster results corresponding to maximum D in Eq. (4), would be near-complete subgraph and the link amount is close to $n(n-1)/2$, where n is the number of member within the cluster.

4.2 Representativeness Metric

After revealing the knowledge structure or exploring topics, overlaps and non-overlaps are easily recognized. And we propose a new metric, representativeness to measure these candidate sentences. As mentioned in Sect. 3, the goal of optimization over criteria is indirectly realized via the trade-off process between overlaps and non-overlaps. Thus representativeness metric should reflect all of sentence salience, length and content redundancy.

To measure sentence salience, LexRank [5] is employed based on the graph model, where eigenvector of the largest eigenvalue is calculated via power method and corresponding eigenvector entries are cast as salience score. And some modifications is applied to LexRank in this paper to make salience score globally meaningful, which is shown in Eq. (5). The basic principle is that due to the effect of normalization of LexRank, the scores of nodes in larger cluster is relatively lower than those in smaller cluster. So the reverse value of highest salience score in a cluster is set as the bench.

$$S_{nc} = \frac{L_c(n)}{L_c(M_c)} \cdot \frac{1}{L_c(M_c)} \tag{5}$$

In Eq. (5), $L_c(n)$ is the LexRank score of node n in cluster c. $L_c(M_c)$ refers the highest LexRank value in cluster c. Set C refers to the clusters node n represent in reality. Then S_{nc} is the global salience score of node n covering cluster c. It is easily recognized that the sentence with higher salience score is likely the core in a large cluster, which means it containing most important topic.

Besides of salience, the factor of summary size should also get involved dynamically. As mentioned above, the goal is indirectly achieved by containing more overlaps, which helps to fulfill the constraint of full topic coverage with less sentences. Then we add coverage bonus metric to reflect overlaps ability to cover topics, which is simply set as the square of the amount of topics the sentence covers. Note that only the sentence with highest salience for every topic would be selected if coverage bonus is simply set as the amount of topics the sentence covers. Because in that case, the sum of the coverage bonus for all selected sentences always equals N since every topic should be covered by single sentence according to the constraint in Eq. (1), which means non-sense in terms of the objective function. In fact with the increase of the power, the ability to cover topics become more and more important. And we simply set square to avoid overwhelming effect of high power. Then representativeness is formally proposed in following Eq. (6),

$$Rep_n = \frac{\sum_c^C \Phi_n(c) S_{nc}}{\Phi_n} + \beta[\sum_c^C \Phi_n(c)]^2 \tag{6}$$

where n is the candidate node and c is the explored topic. $\Phi_n(c)$ equals 1 if sentence n is covering topic c, and 0 otherwise. Generally, there are two parts to form the representativeness. Φ_n refer the amount of topics that sentence n could cover. And the first part of Eq. (6) could be viewed as salience density score of a candidate sentence. And the second part is the corresponding coverage bonus, which is simply set as the square of the amount of topics candidate sentence n really covers. Limited to sentence length, overlaps might have lower salience density than non-overlaps. Nevertheless overlaps would enjoy higher coverage bonus, which helps to compete with non-overlaps in terms of the metric in Eq. (6). Most of time, these two factors are incompatible where broader coverage might be the result of including more low-salience-score overlaps while pursing higher salience density usually means including more non-overlaps with high salience score. β is the coefficient to control this optimization, ranging from 0 to infinite. When β is small, salience density dominates while coverage casts little influence. In this case, the optimization process resembles a salience-based method. As β increases, the topic coverage factor is more and more important, which leads to including more overlaps in the summary.

We also claim that content redundancy is implicitly involved in representativeness metric. When β increases, the performance that the amount of topics a sentence really covers is the key for the metric in Eq. (6). In this case, sentences would be preferred when they can cover multiple topics and more importantly avoid repetitive topic coverage. Since the constraint that a topic should be covered by single sentence, repetitive topic coverage would contribute nothing to representativeness score. Thus, content redundancy is reduced with the increasing β. Conversely when β decrease, salience density of a candidate is the key. In this case, the scenario of repetitive coverage is ignored. And in fact, many sentences contain multiple topics instead of pure one. So this mechanism would definitely increase the possibility of redundant content.

4.3 Overlap-Based Greedy Pick

Since this constrained optimization problem is complicated and NP-hard, we propose Overlap-based Greedy Picking(OGP) algorithm for solution. OGP generates summaries by recursively selecting the candidate sentence with currently highest representativeness scores until achieving full topic coverage. The pseudo code is shown in Algorithm 1.

In each recursion in Algorithm 1, the main purpose is to select the candidate sentence with currently highest representativeness score. Firstly for all available sentences, *GetCurrentAvailableTopics* is the function to find all topics which could be covered by sentence n but not yet covered by other sentences picked before. *GetContainTopics* is the function to calculate the amount of topics sentence n contains. Then the corresponding representativeness score is calculated as Eq. (6). After traversing all sentences, the one with highest score is picked out and included into the summary. Then, *RemovePick* is designed to remove the picked sentence and covering topics out of corresponding sets. At last next recursion begins until all topics are covered.

Although a greedy meghod, we claim that OGP shown in Algorithm 1 is guaranteed near-optimal, above $(1-1/e)$ of the optimal solution. To start with, we show some related theorems. Theorem 1 shows that a function is submodular if it has a diminishing return property. And Theorem 2 shows if the objective function is submodular and monotonic, the performance of the greedy solution is above $(1-1/e)$ of the optimal solution. Next we prove that the designed objective function in Eq. (1) is a monotonic submodular function. Obviously Eq. (1) is monotone increasing, because representativeness score in Eq. (6) is always non-negative, which leads to non-decreasing objective function. Then we prove the submodularity of Eq. (1), which is formally presented in proof. N is the current generated summary containing all previous selected sentences. OF is the objective function in Eq. (1). Then the diminishing property of submodularity in Theorem 1 is directly proved by comparing increasing level of OF when including a sentence into the summary. Note that $\Phi_n^N(c)$ under the smaller background N is not less than $\Phi_n^{N'}(c)$ under a larger N', since some available topics might be covered by others when the amount of sentences in current summary is increased. Finally as OF is monotonic and submodular, the performance of OGP shown in

Algorithm 1 is guaranteed above $(1-1/e)$ of the optimal solution according to Theorem 2.

Theorem 1. *A function $F : 2^V \rightarrow R$ is submodular if for all $s \in V$ and every $S \subseteq S' \subseteq V$, it satisfies $F(S \cup s) - F(S) \geq F(S' \cup s) - F(S')$*

Theorem 2. *Suppose F is monotonic and submodular. Then greedy algorithm gives constant factor approximation: $F(A_{greedy}) \geq (1-1/e) maxF(A)$*

Proof. Submodularity of Objective Function (OF)
Suppose N$\subseteq$ N$'$ and $f(\tilde{N}) = OF(\tilde{N} \cup n) - OF(\tilde{N}) = Rep_n = \frac{\sum_c^C \Phi_n(c) S_{nc}}{\Phi_n} + \beta[\sum_c^C \Phi_n(c)]^2$
$f(N')-f(N) = \frac{\sum_c^C[\Phi_n^{N'}(c)-\Phi_n^N(c)]S_{nc}}{\Phi_n} + \beta\sum_c^C[\Phi_n^{N'}(c)-\Phi_n^N(c)]\sum_c^C[\Phi_n^{N'}(c)+\Phi_n^N(c)]$
Since $\Phi_n(c), \Phi_n, S_{nc} \geq 0$ and $\Phi_n^{N'}(c) \leq \Phi_n^N(c), \forall n$
Then $f(N') - f(N) \leq 0$ and $OF(N' \cup n) - OF(N') \leq OF(N \cup n) - OF(N)$
So the Objective Function or OF is submodular.

Algorithm 1. Overlap-based Greedy Pick

```
INPUT: $C_R$ is the set of topics currently not covered; $N_R$ is the set of overlaps whose
    potential covering topics are not yet completely covered by picked ones.
Function: OGP($C_R$, $N_R$)
if $Length(C_R) = 0$ or $Length(N_R) = 0$ then
    return 0
end if
$Repre \leftarrow 0$
$N_{max} \leftarrow NULL$
for $n$ in $N_R$ do
    $C_r \leftarrow GetCurrentAvailableTopics(C_R, n)$
    $Salience \leftarrow 0$
    $CoverTopics \leftarrow 0$
    for $c$ in $C_r$ do
        $Salience \leftarrow Salience + L_c(O)/L_c(M_c)^2$
        $CoverTopics \leftarrow CoverTopics + 1$
    end for
    $Repre_n \leftarrow Salience/GetContainTopics(n) + \beta * CoverTopics^2$
    if $Repre < Repre_n$ then
        $N_{max} \leftarrow n$
    end if
end for
$C_R$, $N_R \leftarrow RemovePick(C_R, N_R, N_{max})$
return OGP($C_R$, $N_R$) + $Repre$
```

5 Experiment

5.1 DataSets

CL-Scisumm 2014 is the subtrack of TAC 2014 Biomedical Summarization Track. The dataset contains 10 references papers, each of which has up to 10 Citing Papers. Three annotators are employed to manually annotate corresponding citing text spans and reference text spans, which leads to three little different training datasets. To avoid the difference between three annotators, all outcomes are the average of results generated based on these three datasets. Then for this extractive summarization task, the gold summary is generated by manually selecting important citances from the citation context. In this paper, we simply focus on the Task 2, that is generating a faceted summary of up to 250 words based on the reference spans and citing spans. And in order to comply with the length constraint, we simply select top 250 words in case of length exceeding.

5.2 Evaluation Method

Based on the given gold summaries, officially ROUGE-L is employed to measure summaries. ROUGE [13] stands for Recall-Oriented Understudy for Gist Evaluation, which serves as a metric to determine the quality of a summary by comparing it to the ideal one. It is proved highly correlated to human evaluation [13] and chosen to be the standard measure in DUC 2004 summarization tasks. ROUGE-L is a sentence-level metric, focusing on the similarity in terms of longest common subsequence. Three detailed metrics are presented, those are average recall, average precision and average f1-measure. Since the size of a summary would itself cast influence on the precision and recall, average f1-measure value is served as the ultimate criteria.

5.3 Baseline Approaches

In order to verify the performance of the proposed optimization approaches, CLexRank, MEAD and ACL Anthology online toolkit are employed to generate summaries above CL-Scisumm 2014 mentioned in Sect. 5.1. Also the best-recorded system in the competition or MQ is included for comparison.

C-LexRank [29] also employ a graph-based model. It first employs a hierarchial agglomeration algorithm [3] to explore clusters from the graph model. And each cluster of sentences represent a topic. Then it employs LexRank to score sentence within each topic and picks the most salient sentence per topic to form the summary. However, the graph clustering method it employs [3] simply assumes that one sentence contains only one topic and pays no attention to the potential overlap structure.

MEAD [32] is a multi-document summarizer, which generate summaries by picking top-n sentences sorted by scores. Assigned by a feature-based classifier, the score is calculated by linear combination of centroid and position.

ACL Anthology [33] contains all papers published by ACL and other related organizations. It employs string-based heuristics approaches to extract all the citation sentences for an article and generates a citation-based summary, which contains five sentences.

6 Results and Discussion

6.1 Overall Performance

Table 1 shows result comparison between OGP and other baseline approaches. In practice, OGP$_{\beta=0.027}$ is chosen as the representation. And it is intuitive to find that OGP$_{\beta=0.027}$ is better than CLexRank and MEAD in terms of ROUGE-L F-score. Further OGP$_{\beta=0.027}$ is even shorter than both of CLexRank and MEAD, which definitely increases reliability. ACL toolkit is restricted to only five sentences, whose ROUGE-L F-score is low. In order to avoid the influence of summary size, OGP$_{TOP5}$ is to pick top five sentences sorted by representativeness score. And the result of the comparison between OGP$_{TOP5}$ and ACL toolkit shows OGP surpassing ACL default summarization method. As for MQ, all we know is the ROUGE-L F-score, which is still lower than the proposed method. Figure 1 shows the change trend of summary quality generated by OGP with different β. Obviously the above conclusion still works regardless of β.

Table 1. Comparison results of OGP and other baseline approaches.

ROUGE_METHOD		OGP$_{\beta=0.027}$	OGP$_{TOP5}$	CLexRank	MEAD	ACL	MQ
ROUGE-L	AVG_R	0.40982	0.32776	0.37516	0.37466	0.13453	
	AVG_P	0.68594	0.63628	0.63313	0.56310	0.25407	
	AVG_F	0.49795	0.42477	0.45975	0.44146	0.17197	0.260
AVG_SIZE		195.26	119.71	195.93	221.17	115.7	

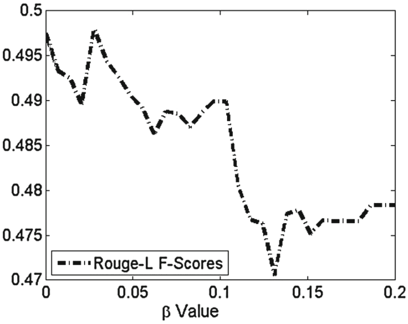
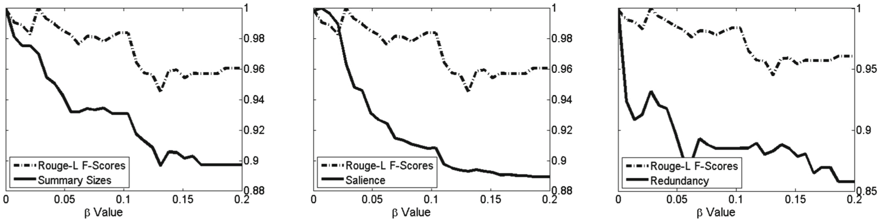


Fig. 1. ROUGE-L F-scores of OGP with various β



(a) (b) (c)

Fig. 2. β's controlling effect over summary size, salience and content redundancy. All of three Figures also reflect the change of ROUGE-L F-score with different β. Figure 2a shows β effect on summary size. Figure 2a shows the influence of β on summary salience score. Figure 2c shows the β effect on content redundancy.

6.2 Effectiveness of β

β is the parameter designed to control the trade-off process between various criteria according to OGP shown in Algorithm 1. When β increases, summary size is reduced since more overlaps are included. Meanwhile content redundancy is also improved since in this case overlaps are preferred when they can avoid topic repetitive coverage. However limited to sentence length, more overlaps might reduce summary salience. To verify the conclusion, summaries are generated by OGP with different β and the result is shown in Fig. 2. In order to make the result more intuitive, all data shown in Fig. 2 is normalized by dividing maximum value, which makes y value of each figure range from 0 to 1. Figure 2a specifically shows the change of summary size with different β. The result turns out that summary size is reduced up to 10% with increasing β. Figure 2b shows how salience score shown in Eq. (5) changes with different β. From the Fig. 2b, salience is nearly monotone decreasing and reduced up to 11%. Figure 2c shows the change of content redundancy with different β. Content redundancy is defined as number of times that topics are repetitive covered. Note that ideally every topic is covered once where redundancy should be zero. Thus redundancy should be normalized via a two-step process. Firstly redundancy is subtracted by the amount of topics. Then normalized redundancy is calculated by dividing current maximum redundancy value. According to the results from Fig. 2c, content redundancy generally decreases with the increasing β, whose rate is up to 14%. Generally, the results shown in Fig. 2 comply with the conclusion of β effect mentioned in Sect. 4.1.

ROUGE-L F-score is used to measure the summary quality. Although fluctuations exists and the rate is relatively small, summary quality generally decreases with increasing β from Fig. 1. Note that longer summary size is definitely beneficial for improving summary quality, since topics might be explained better with more sentences. Thus the scenario of summary quality decreasing could be easily explained by the synthetic effect by summary size, salience and content redundancy. Since the sum of the drop rate of both summary salience and salience is greater than the improve rate of redundancy, the synthetic effect reflected to the general summary quality becomes negative with increasing β, which could be easily recognized from Fig. 2. Also since the gap between these rate is small and the gap is not always negative, the drop of ROUGE-L F-score is slow and unstable.

7 Conclusion and Future Work

In this paper, we mainly propose a new optimization methodology to generate citation-based summarization. Criteria such as summary size, salience and content redundancy is indirectly involved via the optimization between overlaps and non-overlaps, which successfully solve the problem of inconsistent metrics. To solve the optimization problem, a performance-guaranteed greedy algorithm OGP is proposed. Further via a single parameter β, balance status of criteria is effectively controlled.

There are still several tasks for future work. Firstly, the value of β in the experiment is limited to a certain range in advance. And the task to find proper β could be studied in future. Secondly and most importantly, the change of summary quality is strongly related to the synthetic effect of summary size, salience and content redundancy. So the task to explore such relationship is extremely important for better understanding the contribution of each criterion.

References

1. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature, pp. 761–764 (2010)
2. Carbonell, J., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and development in information retrieval, pp. 335–336 (1998)
3. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Phys. Rev. E **70**, 1–6 (2004)
4. Erkan, G., Radev, D.R.: Lexpagerank: Prestige in multi-document text summarization. In: Conference on Empirical Methods in Natural Language Processing, pp. 365–371 (2004)
5. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004)
6. Filatova, E., Hatzivassiloglou, V.: A formal model for information selection in multi-sentence text extraction. In: Proceedings of the 20th International Conference on Computational Linguistics, pp. 397–403 (2004)
7. Fung, P., Ngai, G., Cheung, C.S.: Combining optimal clustering and hidden Markov models for extractive summarization. In: Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering, pp. 21–28 (2003)
8. Harabagiu, S., Lacatusu, F.: Topic themes for multi-document summarization. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 202–209 (2005)

9. Hardy, H., Shimizu, N., Strzalkowski, T., Ting, L., Zhang, X., Wise, G.B.: Cross-document summarization by concept classification. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 121–128 (2002)
10. Hirao, T., Yoshida, Y., Nishino, M., Yasuda, N., Nagata, M.: Single-document summarization as a tree knapsack problem. In: Conference on Empirical Methods in Natural Language Processing, pp. 1515–1520 (2013)
11. Jaccard, P.: Etude comparative de la distribution florale dans une portion des Alpes et du Jura. Impr. Corbaz (1901)
12. Kaplan, D., Iida, R., Tokunaga, T.: Automatic extraction of citation contexts for research paper summarization: a coreference-chain based approach. In: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, pp. 88–95 (2009)
13. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74–81 (2004)
14. Lin, H., Bilmes, J., Xie, S.: Graph-based submodular selection for extractive summarization. In: IEEE Workshop on Automatic Speech Recognition & Understanding, ASRU 2009, pp. 381–386 (2009)
15. Lin, H., Bilmes, J.: Multi-document summarization via budgeted maximization of submodular functions. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 912–920 (2010)
16. Lin, H., Bilmes, J.: A class of submodular functions for document summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 510–520 (2011)
17. Lin, H., Bilmes, J.: Learning mixtures of submodular shells with application to document summarization. In: Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, pp. 479–490 (2012)
18. McDonald, R.: A study of global inference algorithms in multi-document summarization. In: Amati, G., Carpineto, C., Romano, G. (eds.) ECIR 2007. LNCS, vol. 4425, pp. 557–564. Springer, Heidelberg (2007). doi:10.1007/978-3-540-71496-5_51
19. McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., Eskin, E.: Towards multidocument summarization by reformulation: progress and prospects. In: Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence, pp. 453–460 (1999)
20. Mei, Q., Guo, J., Radev, D.: Divrank: the interplay of prestige and diversity in information networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1009–1018 (2010)
21. Mei, Q., Zhai, C.: Generating impact-based summaries for scientific literature. In: Proceedings of the Meeting of the Association for Computational Linguistics, pp. 816–824 (2008)
22. Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., Zajic, D.: Using citations to generate surveys of scientific paradigms. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 584–592 (2009)
23. Morita, H., Sasano, R., Takamura, H., Okumura, M.: Subtree extractive summarization via submodular maximization. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 1023–1032 (2013)

24. Nakov, P.I., Schwartz, A.S., Hearst, M.: Citances: citation sentences for semantic analysis of bioscience text. In: Proceedings of the SIGIR 2004 workshop on Search and Discovery in Bioinformatics, pp. 81–88 (2004)
25. Nanba, H., Okumura, M.: Towards multi-paper summarization using reference information. In: International Joint Conference on Artificial Intelligence, pp. 926–931 (1999)
26. Nishikawa, H., Hirao, T., Makino, T., Matsuo, Y.: Text summarization model based on redundancy-constrained knapsack problem. In: Proceedings of COLING 2012: Posters, pp. 893–902 (2012)
27. Parveen, D., Mesgar, M., Strube, M.: Generating coherent summaries of scientific articles using coherence patterns. In: Conference on Empirical Methods in Natural Language Processing, pp. 772–783 (2016)
28. Parveen, D., Ramsl, H.M., Strube, M.: Topical coherence for graph-based extractive summarization, pp. 1949–1954 (2015)
29. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proceedings of the 22nd International Conference on Computational Linguistics, pp. 689–696 (2008)
30. Qazvinian, V., Radev, D.R.: Identifying non-explicit citing sentences for citation-based summarization. In: Proceedings of the 48th annual meeting of the association for computational linguistics, pp. 555–564 (2010)
31. Qian, X., Liu, Y.: Fast joint compression and summarization via graph cuts. In: Conference on Empirical Methods in Natural Language Processing, pp. 1492–1502 (2013)
32. Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., Liu, D., et al.: Mead-a platform for multidocument multilingual text summarization (2004)
33. Radev, D.R., Muthukrishnan, P., Qazvinian, V., Abu-Jbara, A.: The ACL anthology network corpus. Lang. Resour. Eval. **47**, 919–944 (2013)
34. Shen, C., Li, T.: Multi-document summarization via the minimum dominating set. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 984–992 (2010)
35. Siddharthan, A., Nenkova, A., McKeown, K.: Syntactic simplification for improving content selection in multi-document summarization. In: Proceedings of the 20th international conference on Computational Linguistics, pp. 896–902 (2004)
36. Skabar, A., Abdalgader, K.: Clustering sentence-level text using a novel fuzzy relational clustering algorithm. IEEE Trans. Knowl. Data Eng. **25**, 62–75 (2013)
37. Takamura, H., Okumura, M.: Text summarization model based on maximum coverage problem and its variant. In: Conference of the European Chapter of the Association for Computational Linguistics, pp. 505–513 (2009)
38. Vigneshwaran, L.J.K.P.M., Sharma, M.V.V.D.M.: Non-decreasing sub-modular function for comprehensible summarization. In: Proceedings of NAACL-HLT, pp. 94–101 (2016)