Optimizing an Approximation of ROUGE – a Problem-Reduction Approach to Extractive Multi-Document Summarization

Maxime Peyrard and **Judith Eckle-Kohler**
Research Training Group AIPHES and UKP Lab
Computer Science Department, Technische Universität Darmstadt
www.aiphes.tu-darmstadt.de, www.ukp.tu-darmstadt.de

Abstract

This paper presents a problem-reduction approach to extractive multi-document summarization: we propose a reduction to the problem of scoring individual sentences with their ROUGE scores based on supervised learning. For the summarization, we solve an optimization problem where the ROUGE score of the selected summary sentences is maximized. To this end, we derive an approximation of the ROUGE-N score of a set of sentences, and define a principled discrete optimization problem for sentence selection. Mathematical and empirical evidence suggests that the sentence selection step is solved almost exactly, thus reducing the problem to the sentence scoring task. We perform a detailed experimental evaluation on two DUC datasets to demonstrate the validity of our approach.

1 Introduction

Multi-document summarization (MDS) is the task of constructing a summary from a topically related document collection. This paper focuses on the variant of extractive and generic MDS, which has been studied in detail for the news domain using available benchmark datasets from the Document Understanding Conference (DUC) (Over et al., 2007).

Extractive MDS can be cast as a budgeted subset selection problem (McDonald, 2007; Lin and Bilmes, 2011) where the document collection is considered as a set of sentences and the task is to select a subset of the sentences under a length constraint. State-of-the-art and recent works in extractive MDS solve this discrete optimization problem using integer linear programming (ILP) or submodular function maximization (Gillick and Favre, 2009; Mogren et al., 2015; Li et al., 2013b; Kulesza and Taskar, 2012; Hong and Nenkova, 2014). The objective function that is maximized in the optimization step varies considerably in previous work. For instance, Yih et al. (2007) maximize the number of informative words, Gillick and Favre (2009) the coverage of particular concepts, and others maximize a notion of "summary worthiness", while minimizing summary redundancy (Lin and Bilmes, 2011; Kågebäck et al., 2014).

There are also multiple approaches which maximize the evaluation metric for system summaries itself based on supervised Machine Learning (ML). System summaries are commonly evaluated using ROUGE (Lin, 2004), a recall oriented metric that measures the n-gram overlap between a system summary and a set of human-written reference summaries.

The benchmark datasets for MDS can be employed in two different ways for supervised learning of ROUGE scores: either by training a model that assigns ROUGE scores to individual textual units (e.g., sentences), or by performing structured output learning and directly maximizing the ROUGE scores of the created summaries (Nishikawa et al., 2014; Takamura and Okumura, 2010; Sipos et al., 2012). The latter approach suffers both from the limited amount of training data and from the higher complexity of the machine learning models.

In contrast, supervised learning of ROUGE scores for individual sentences can be performed with simple regression models using hundreds of sentences as training instances, taken from a single pair of documents and reference summaries. Extractive MDS can leverage the ROUGE scores of individual sentences in various ways, in particular, as part of an optimization step. In our work, we follow the previously successful approaches to

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1825–1836,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

extractive MDS using discrete optimization, and make the following contributions:

We provide a theoretical justification and empirical validation for using ROUGE scores of individual sentences as an optimization objective. Assuming that ROUGE scores of individual sentences have been estimated by a supervised learner, we derive an approximation of the ROUGE-N score for a set of sentences from the ROUGE-N scores of the individual sentences in the general case of $N >= 1$.

We use our approximation to define a mathematically principled discrete optimization problem for sentence selection. We empirically evaluate our framework on two DUC datasets, demonstrating the validity of our approximation, as well as its ability to achieve competitive ROUGE scores in comparison to several strong baselines.

Most importantly, the resulting framework reduces the MDS task to the problem of scoring individual sentences with their ROUGE scores. The overall summarization task is converted to two sequential tasks: (i) scoring single sentences, and (ii) selecting summary sentences by solving an optimization problem where the ROUGE score of the selected sentences is maximized.

The optimization objective we propose almost exactly solves (ii), which we justify by providing both mathematical and empirical evidence. Hence, solving the whole problem of MDS is reduced to solving (i).

The rest of this paper is structured as follows: in Section 2, we discuss related work. Section 3 presents our subset selection framework consisting of an approximation of the ROUGE score of a set of sentences, and a mathematically principled discrete optimization problem for sentence selection. We evaluate our framework in Section 4 and discuss the results in Section 5. Section 6 concludes.

2 Related Work

Related to our approach is previous work in extractive MDS that (i) casts the summarization problem as budgeted subset selection, and (ii) employs supervised learning on MDS datasets to learn a scoring function for textual units.

Budgeted Subset Selection Extractive MDS can be formulated as the problem of selecting a subset of textual units from a document collection such that the overall score of the created summary is maximal and a given length constraint is observed. The selection of textual units for the summary relies on their individual scores, assigned by a scoring function which represents aspects of their relevance for a summary. Often, sentences are considered as textual units.

Simultaneously maximizing the relevance scores of the selected units and minimizing their pairwise redundancy given a length constraint is a global inference problem which can be solved using ILP (McDonald, 2007). Several state-of-the-art results in MDS have been obtained by using ILP to maximize the number of relevant concepts in the created summary while minimizing the pairwise similarity between the selected sentences (Gillick and Favre, 2009; Boudin et al., 2015; Woodsend and Lapata, 2012).

Another way to formulate the problem of finding the best subset of textual units is to maximize a submodular function. Maximizing submodular functions is a general technique that uses a greedy optimization algorithm with a mathematical guarantee on optimality (Nemhauser and Wolsey, 1978). Performing summarization in the framework of submodularity is natural because summaries try to maximize the coverage of relevant units while minimizing redundancy (Lin and Bilmes, 2011). However, several different coverage and redundancy functions have been proposed (Lin and Bilmes, 2011; Kågebäck et al., 2014; Yin and Pei, 2015) recently, and there is not yet a clear consensus on which coverage function to maximize.

Supervised Learning Supervised learning using datasets with reference summaries has already been employed in early work on summarization to classify sentences as summary-worthy or not (Kupiec et al., 1995; Aone et al., 1995).

Learning a scoring function for various kinds of textual units has become especially popular in the context of global optimization: scores of textual units, learned from data, are fed into an ILP problem solver to find the subset of sentences with maximal overall score. For example, Yih et al. (2007) score each word in the document cluster based on frequency and position, Li et al. (2013b) learn bigram frequency in the reference summaries, and Hong and Nenkova (2014) learn word importance from a rich set of features.

Closely related to our work are summarization approaches that include a supervised component which assigns ROUGE scores to individual sentences. For example, Ng et al. (2012), Li et al. (2013a) and Li et al. (2015) all use a regression model to learn ROUGE-2 scores for individual sentences, but use it in different ways for the summarization. While Ng et al. (2012) use the ROUGE scores of sentences in combination with the Maximal Marginal Relevance algorithm as a baseline approach, Li et al. (2013a) use the scores to select the top-ranked sentences for sentence compression and subsequent summarization. Li et al. (2015), in contrast, use the ROUGE scores to re-rank a set of sentences that are output by an optimization step.

While learning ROUGE scores of textual units is widely used in summarization systems, the theoretical background on why this is useful has not been well studied yet. In our work, we present the mathematical and empirical justification for this common practice. In the next section, we start with the mathematical justification.

3 Content Selection Framework

3.1 Approximation of ROUGE-N

Notation: Let $S = \{s_i | i \leq m\}$ be a set of m sentences which constitute a system summary. We use $\rho_N(S)$ or simply $\rho(S)$ to denote the ROUGE-N score of S. ROUGE-N evaluates the n-gram overlap between S and a set of reference summaries (Lin, 2004). Let S^* denote the reference summary and R_N the number of n-gram tokens in S^*. R_N is a function of the summary length in words, in particular, R_1 is the target size of the summary in words. Finally, let $F_S(g)$ denote the number of times the n-gram type g occurs in S. For a single reference summary, ROUGE-N is computed as follows:

$$\rho(S) = \frac{1}{R_N} \sum_{g \in S^*} min(F_S(g), F_{S^*}(g)) \qquad (1)$$

For compactness, we use the following notation for any set of sentences X:

$$C_{X,S^*}(g) = min(F_X(g), F_{S^*}(g)) \qquad (2)$$

$C_{X,S^*}(g)$ can be understood as the contribution of the n-gram g.

ROUGE-N for a Pair of Sentences: Using this notation, the ROUGE-N score of a set of two sentences a and b can be written as:

$$\rho(a \cup b) = \frac{1}{R_N} \sum_{g \in S^*} min(C_{a \cup b, S^*}(g), F_{S^*}(g)) \qquad (3)$$

We observe that $\rho(a \cup b)$ can be expressed as a function of the individual scores $\rho(a)$ and $\rho(b)$:

$$\rho(a \cup b) = \rho(a) + \rho(b) - \epsilon(a \cap b) \qquad (4)$$

where $\epsilon(a \cap b)$ is an error correction term that discards overcounted n-grams from the sum of $\rho(a)$ and $\rho(b)$:

$$\epsilon(a \cap b) = \frac{1}{R_N} \sum_{g \in S^*} max(C_{a,S^*}(g) + C_{b,S^*}(g) - F_{S^*}(g), 0) \qquad (5)$$

A proof that this error correction is correct is given in appendix A.1.

General Formulation of ROUGE-N: We can extend the previous formulation of ρ to sets of arbitrary cardinality using recursion. If $\rho(S)$ is given for a set of sentences S, and a is a sentence then:

$$\rho(S \cup a) = \rho(S) + \rho(a) - \epsilon(S \cap a) \qquad (6)$$

We prove in appendix A.1 that this formula is the ROUGE-N score of $S \cup a$.

Another way to obtain ρ for an arbitrary set S is to adapt the principle of inclusion-exclusion:

$$\rho(S) = \sum_{i=1}^{m} \rho(s_i) + \sum_{k=2}^{m} (-1)^{k+1} \Big(\sum_{1 \leq i_1 \leq \cdots \leq i_k \leq m} \epsilon^{(k)}(s_{i_1} \cap \cdots \cap s_{i_k}) \Big) \qquad (7)$$

This formula can be understood as adding up scores of individual sentences, but n-grams appearing in the intersection of two sentences might be overcounted. $\epsilon^{(2)}$ is used to account for these n-grams. But now, n-grams in the intersection of three sentences might be undercounted and $\epsilon^{(3)}$ is used to correct this. Each $\epsilon^{(k)}$ contributes to improving the accuracy by refining the errors made by $\epsilon^{(k-1)}$ for the n-grams appearing in the intersection of k sentences. When $k = |S|$, $\rho(S)$ is exactly the ROUGE-N of S. A rigorous proof and details about $\epsilon^{(k)}$ are provided in appendix A.2.

Approximation of ROUGE-N for a Pair of Sentences: To find a valid approximation of ρ as defined in (7), we first consider the $\rho(a \cup b)$ from equation (3) and then extend it to the general case. When maximizing ρ, scores for sentences are assumed to be given (e.g., estimated by a ML component). We still need to estimate $\epsilon(a \cap b)$, which means, according to (5), to estimate:

$$\sum_{g \in S^*} max(C_{a,S^*}(g) + C_{b,S^*}(g) - F_{S^*}(g), 0) \quad (8)$$

At inference time, neither S^* (the reference summary) nor F_{S^*} (number of occurrences of n-grams in the reference summary) is known.

At this point, we can observe that, similar as for sentence scoring, ϵ can be estimated via a supervised ML component. Such an ML model can easily be trained on the intersections of all sentence pairs in a given training dataset. Hence, we can assume that both the scores for individual sentences and the ϵ are learned empirically from data using ML. As a result, we have pushed all estimation steps into supervised ML components, which leaves the subset selection step fully principled.

However, we found in our experiments that even a simple heuristic yields a decent approximation of ϵ. The heuristic uses the frequency $freq(g)$ of an n-gram g observed in the source documents:

$$\sum_{g \in S^*} max(C_{a,S^*}(g) + C_{b,S^*}(g) - F_{S^*}(g), 0) \approx \sum_{g \in a \cap b} \mathbb{1}[freq(g) \geq \alpha] \quad (9)$$

The threshold α tells us which n-grams are likely to appear in the reference summary, and it is determined by grid-search on the training set. This is penalizing n-grams which appear twice and are likely to occur in the summary. It can be understood as a way of limiting redundancy. In practice, we used $\alpha = 0.3$. However, we experimented with various values of the hyper-parameter α and found that its value has no significant impact as long as it is fairly small (< 0.5). Higher values will ignore too many redundant n-grams and the summary will have a high redundancy.

R_N is known since it is simply the number of n-gram tokens in the summaries. We end up with the following approximation for the pairwise case:

$$\tilde{\rho}(a \cup b) = \rho(a) + \rho(b) - \tilde{\epsilon}(a \cup b), \text{ where}$$
$$\tilde{\epsilon}(a \cup b) = \frac{1}{R_N} \sum_{g \in a \cap b} \mathbb{1}[freq(g) \geq \alpha] \quad (10)$$

General Approximation of ROUGE-N: Now, we can approximate $\rho(S)$ for the general case defined by equation (7). We recall that $\rho(S)$ contains the sum of $\rho(s_i)$, the pairwise error terms $\epsilon^{(2)}(s_i \cap s_j)$, the error terms of three sentences $\epsilon^{(3)}$ and so on.

We can restrict ourselves to the individual sentences and the pairwise error corrections. Indeed, the intersection between more than two sentences is often empty, and accounting for it does not improve the accuracy significantly, but greatly increases the computational cost.

A formulation of ϵ in the case of two sentences has already been defined in (10). Thus, we have an approximation of the ROUGE-N function for any set of sentences that can be computed at inference time:

$$\tilde{\rho}(S) = \sum_{i=1}^{n} \rho(s_i) - \sum_{a,b \in S, a \neq b} \tilde{\epsilon}(a \cap b) \quad (11)$$

We empirically checked the validity of this approximation. For this, we sampled 1000 sets of sentences from source documents of DUC-2003 (sets of 2 to 5 sentences) and compared their $\tilde{\rho}$ score to the real ROUGE-N. We observe a pearson's r correlation ≥ 0.97, which validates $\tilde{\rho}$.

3.2 Discrete Optimization

$\tilde{\rho}$ from equation (11) defines a set function that scores a set of sentences. The task of summarization is now to select the set S^* with maximal $\tilde{\rho}(S^*)$ under a length constraint.

Submodularity: A submodular function is a set function obeying the diminishing returns property: $\forall S \subseteq T$ and a sentence a: $F(S \cup a) - F(S) \geq F(T \cup a) - F(T)$. Submodular functions are convenient because maximization under constraints can be done greedily with a guarantee of the optimality of the solution (Nemhauser et al., 1978).

It has been shown that ROUGE-N is submodular (Lin and Bilmes, 2011) and it is easy to verify that $\tilde{\rho}$ is submodular as well (the proof is given in the supplemental material).

We can therefore apply the greedy maximization algorithm to find a good set of sentences. This has the advantage of being straightforward and fast, however it does not necessarily find the optimal solution.

ILP: A common way to solve a discrete optimization problem is to formulate it as an ILP. It maximizes (or minimizes) a linear objective function with some linear constraints where the variables are integers. ILP has been well studied and existing tools can efficiently retrieve the exact solution of an ILP problem.

We observe that it is possible to formulate the maximization of $\tilde{\rho}(S)$ as an ILP. Let x be the binary vector whose i-th entry indicates whether sentence i is in the summary or not, $\tilde{\rho}(s_i)$ the scores of sentences, and K the length constraint. We pre-compute the symmetric matrix $\tilde{P}$ where $\tilde{P}_{i,j} = \tilde{\epsilon}(s_i \cap s_j)$ and solve the following ILP:

$$max(\sum_{i=1}^{n} x_i * \tilde{\rho}(s_i) - d\frac{1}{R}\sum_{i \geq j} \alpha_{i,j} * \tilde{P}i,j)$$
$$\textstyle\sum_{i=1}^{n} x_i * len(s_i) \leq K$$
$$\forall(i,j), \alpha_{i,j} - x_i \leq 0$$
$$\forall(i,j), \alpha_{i,j} - x_j \leq 0$$
$$\forall(i,j), x_i + x_j - \alpha_{i,j} \leq 1$$

d is a damping factor that allows to account for approximation errors. When $d = 0$, the problem becomes the maximization of “summary worthiness” under a length constraint, with “summary worthiness” being defined by $\rho(s_i)$.

In practice, we used a value $d = 0.9$ because we observed that the learner tends to slightly overestimate the ROUGE-N scores of sentences. The mathematical derivation implies $d = 1$, however we can easily adjust for shifts in average scores of sentences from the estimation step by adjusting d. Another option would be to post-process the scores after the estimation step to fix the average and let $d = 1$ in the optimization step. Indeed, if d moves away from 1, we move away from the mathematical framework of ROUGE-N maximization.

If $d \neq 0$, it seems intuitive to interpret the second term as minimizing the summary redundancy, which is in accordance to previous works.

However, in our framework, this term has a precise interpretation:

it maximizes ROUGE-N scores up to the second order of precision, and the ROUGE-N formula itself already induces a notion of “summary worthiness” and redundancy, which we can empirically infer from data via supervised ML for sentence scoring, and a simple heuristic for sentence intersections.

4 Evaluation

We perform three kinds of experiments in order to empirically evaluate our framework: first, we show that our proposed approximation is valid, then we analyze a basic supervised sentence scoring component, and finally we perform an extrinsic evaluation on end-to-end extractive MDS.

In our experiments, we use the DUC datasets from 2002 and 2003 (DUC-02 and DUC-03). We use the variants of ROUGE identified by Owczarzak et al. (2012) as strongly correlating with human evaluation methods: ROUGE-2 recall with stemming and stopwords not removed (giving the best agreement with human evaluation), and ROUGE-1 recall (as the measure with the highest ability to identify the better summary in a pair of system summaries). For DUC-03, summaries are truncated to 100 words, and to 200 words for DUC-02. [1] The truncation is done automatically by ROUGE. [2]

4.1 Framework Validity

Given that sentences receive scores close to their individual ROUGE-N, we presented a function that approximates the ROUGE-N of sets of these sentences and proposed an optimization to find the best scoring set under a length constraint.

To validate our framework empirically, we consider its upper-bound, which is obtained when our ILP/submodular optimizations use the real ROUGE-N scores of the individual sentences, calculated based on the reference summaries. We compare this upper bound to a greedy approach, which simply adds the best scoring sentences one by one to the subset until the length limit is reached, and to the real upper bound for extractive summarization which is determined by solving a maximum coverage problem for n-grams from the reference summary (as it was done by Takamura and Okumura (2010)).

Table 1 shows the results. We observe that ILP-R produces scores close to the reference, thus reducing the problem of extractive summarization to the task of sentence scoring, because the perfect scores induced near perfect extracted summaries in this framework. SBL-R seems less promising than ILP-R because it greedily maximizes a function which ILP-R exactly maximizes. Therefore, we continue our experiments in the following sec-

[1] In the official DUC-03 competitions, summaries of length 665 bytes were expected. Systems could produce different numbers of words. The variation in length has a noticeable impact on ROUGE recall scores.

[2] ROUGE-1.5.5 with the parameters: -n 2 -m -a -l 100 -x -c 95 -r 1000 -f A -p 0.5 -t 0. The length parameter becomes -l 200 for DUC-02.

tions with ILP-R only. However, SBL-R offers a nice trade-off between performance and computation cost. The greedy optimization of SBL-R is noticeably faster than ILP-R.

	DUC-02		DUC-03	
	R1	R2	R1	R2
Greedy	0.597	0.414	0.391	0.148
SBL-R	0.630	0.484	0.424	0.160
ILP-R	0.644	0.495	0.447	0.178
Upper Bound	0.648	0.497	0.452	0.181

Table 1: Upper bound of our framework compared to extractive upper bound.

In practice, the learner will not produce perfect scores. We experimentally validated that with learned scores converging to true scores, the extracted summary converges to the best extractive summary (w.r.t to ROUGE-N). To this end, we simulated approximate learners by artificially randomizing the true scores to end up with lists having various correlations with the true scores. We fed these scores to ILP-R and computed the ROUGE-1 of the generated summaries for an example topic from DUC-2003. Figure 1 displays the expected ROUGE-1 versus the performance of the artificial learner (correlation with true scores of sentences). We observe that, as the learner improves, the generated summaries approach the best ROUGE scoring summary.

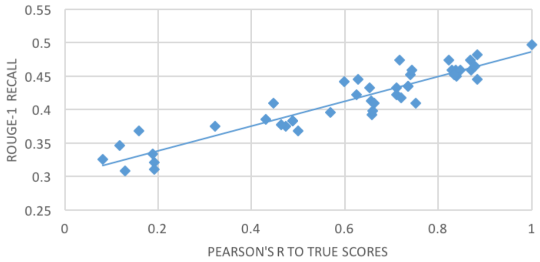

Figure 1: ROUGE-1 of summary against sentence scores correlation with true ROUGE-1 scores of sentences (d30003t from DUC-2003).

4.2 Sentence Scoring

Now we look at the supervised learning component which learns ROUGE-N scores for individual sentences. We know that we can achieve an overall summary ROUGE-N score close to the upper bound, if a learner would be able to learn the scores perfectly. For better understanding the difficulty of the task of sentence scoring, we look at the correlation of the scores produced by a basic learner and the true scores given in a reference dataset.

Model and Features From an existing summarization dataset (e.g. a DUC dataset), a training set can straightforwardly be extracted by annotating each sentence in the source documents with its ROUGE-N score. For each topic in the dataset, this yields a list of sentences and their target score.

To support the claim that learning ROUGE scores for individual sentences is easier than solving the whole summarization task, it is sufficient to choose a basic learner with simple features and little in-domain training data (models are trained on one DUC dataset and evaluated on another). Specifically, we employ a support vector regression (SVR).[3] We use only classical surface-level features to represent sentences (position, length, overlap with title) and combine them with frequency features. The latter include TF*IDF weighting of the terms (similar to Luhn (1958)), the sum of the frequency of the bi-grams in the sentence, as well as the sum of the document frequency (number of source documents in which the n-grams appear) of the terms and bi-grams in a sentence.

We trained two models, R1 and R2 on DUC-02 and DUC-03. For R1, the target score is the ROUGE-1 recall, while R2 learns ROUGE-2 recall.

Correlation Analysis We evaluated our sentence scoring models R1 and R2 by calculating the correlation of the scores produced by R1 and R2 and the true scores given in the DUC-03 data. We compare both models to the true ROUGE-1 and ROUGE-2 scores. In addition, we calculated the correlation of the TF*IDF and LexRank scores, in order to understand how well they would fit into our framework (TF*IDF and LexRank are described in section 4.3).

The results are displayed in Table 2. Even with a basic learner it is possible to learn scores that correlate well with the true ROUGE-N scores, which supports the claim that it is easier to learn scores for individual sentences than to solve the whole problem of summarization. This finding strongly supports our proposed reduction of the extractive MDS problem to the task of learning

[3] We use the implementation in scikit-learn (Pedregosa et al., 2011).

	with ROUGE-1			with ROUGE-2		
	Pearson's r	Kendall's tau	nDCG@15	Pearson's r	Kendall's tau	nDCG@15
TF*IDF	0.923	0.788	0.916	0.607	0.512	0.580
LexRank	0.210	0.120	0.534	0.286	0.178	0.379
model R1	**0.940**	**0.813**	**0.951**	0.653	0.545	0.693
model R2	0.729	0.496	0.891	**0.743**	**0.576**	**0.752**

Table 2: Correlation of different kinds of sentence scores and their true ROUGE-1 and ROUGE-2 scores.

scores for individual sentences, which correlate well with their true ROUGE-N scores.

We observe that TF*IDF correlates surprisingly well with the ROUGE-1 score, which indicates that we can expect a significant performance gain when feeding TF*IDF scores to our optimization framework. LexRank, on the other hand, orders sentences according to their centrality and does not look at individual sentences. Accordingly, we observe a low correlation with the true ROUGE-N scores, and thus LexRank may not benefit from the optimization (which we confirmed in our experiments).

Finally, we observe that there is significant room for improvement regarding ROUGE-2, as well as for Kendall's tau in ROUGE-1 where a more sophisticated learner could produce scores that correlate better with the true scores. The higher the correlation of the sentence scores assigned by a learner and the true scores, the better the summary produced by the subsequent subset selection.

4.3 End-to-End Evaluation

In our end-to-end evaluation on extractive MDS, we use the following baselines for comparison:

- **TF*IDF weighting**: This simple heuristic was introduced by Luhn (1958). Each sentence receives a score from the TF*IDF of its terms. We trained IDFs (Inverse Document Frequencies) on a background corpus [4] to improve the original algorithm.

- **LexRank**: Among other graph-based approaches to summarization (Mani and Bloedorn, 1997; Radev et al., 2000; Mihalcea, 2004), LexRank (Erkan and Radev, 2004) has become the most popular one. A similarity graph $G(V, E)$ is constructed where V is the set of sentences and an edge e_{ij} is drawn between sentences v_i and v_j if and only if the cosine similarity between them is above a given threshold. Sentences are scored according to their PageRank score in G. For our experiments, we use the implementation available in the sumy package.[5]

- **ICSI**: ICSI is a recent system that has been identified as one of the state-of-the-art systems by Hong et al. (2014). It is a global linear optimization framework that extracts a summary by solving a maximum coverage problem considering the most important concepts in the source documents. Concepts are identified as bi-grams and their importance is estimated via their frequency in the source documents. Boudin et al. (2015) released a Python implementation (ICSI sume) that we use in our experiments.

- **SFOUR**: SFOUR is a structured prediction approach that trains an end-to-end system with a large-margin method to optimize a convex relaxation of ROUGE (Sipos et al., 2012). We use the publicly available implementation. [6]

As described in the previous section, two models are trained: R1 and R2. We evaluate both of them in the end-to-end setup with and without our optimization. In the greedy version, sentences are added as long as the summary length is valid.

We apply the optimization for sentence scoring models trained on ROUGE-1 and ROUGE-2 as well. The scoring models are trained on one dataset and evaluated on the other. For the ILP optimization, the damping factor can vary and leads to different performance. We report the best results among few variations. In order to speed-up the ILP step, we propose to limit the search space by only looking at the top K sentences[7] (hence

[4] We used DBpedia long abstract: http://wiki.dbpedia.org/Downloads2015-04.

[5] https://github.com/miso-belica/sumy

[6] http://www.cs.cornell.edu/˜rs/sfour/

[7] We used K=50 and observed that a range from K=25 to K=70 yields a good trade-off between computation cost and performance.

the importance of learning a correct ordering as well, like Kendall's tau). This results in a massive speed-up and can even lead to better results as it prunes parts of the noise. Finally, we perform significance testing with the t-test to compare differences between two means.[8]

	DUC-02		DUC-03	
	R1	R2	R1	R2
TFIDF	0.403	0.120	0.322	0.066
LexRank	0.446	0.158	0.354	0.077
ICSI	0.445	0.155	0.375	0.094
SFOUR	0.442	0.181	0.365	0.087
Greedy-R1	0.480	0.115	0.353	0.084
Greedy-R2	0.499	0.132	0.369	0.093
TFIDF+ILP	0.415	0.135	0.335	0.075
R1+ILP	0.509	0.187	0.378	0.101
R2+ILP	**0.516***	**0.192***	**0.379**	**0.102**

Table 3: Impact of the optimization step on sentence subset selection.

Results Table 3 shows the results. The proposed optimization significantly and systematically improves TF*IDF performance as we expected from our analysis in the previous section. This result suggests that using only a frequency signal in source documents is enough to get high scoring summaries, which supports the common belief that frequency is one of the most useful features for generic news summarization. It also aligns well with the strong performance of ICSI, which combines an ILP step with frequency information as well.

The optimization also significantly and systematically improves upon the greedy approach combined with our scoring models. Combining a SVR learner (SVR-1 and SVR-2) and our ILP-R produces results on par with ICSI and sometimes significantly better. SFOUR maximizes ROUGE in an end-to-end fashion, but is outperformed by our framework when using the same training data. The framework is able to reach a competitive performance even with a basic learner. These results again suggest that investigating better learners for sentence scoring might be promising in order to improve the quality of the summaries.

We observe that the model trained on ROUGE-2 is performing better than the model trained on ROUGE-1, although learning the ROUGE-2 scores seems to be harder than learning ROUGE-1 scores (as shown in table 2). However, errors and approximations propagate less easily in ROUGE-2, because the number of bi-grams in the intersection of two given sentences is far less. Hence we conclude that learning ROUGE-2 scores should be put into the focus of future work on improving sentence scoring.

5 Discussion

This section discusses our contributions in a broader context.

ROUGE Our subset selection framework performs the task of *content selection*, selecting an unordered set of textual units (sentences for now) for a system summary. The re-ordering of the sentences is left to a subsequent processing step, which accounts for aspects of discourse coherence and readability.

While we justified our choice of ROUGE-1 recall and ROUGE-2 recall as optimization objectives by their strong correlation with human evaluation methods, ROUGE-N has also various drawbacks. In particular, it does not take into account the overall discourse coherence of a system summary (see the supplemental material for examples of summaries generated by our framework).

From a broader perspective, systems that have high ROUGE scores can only be as good as ROUGE is, as a proxy for summary quality. However, as long as systems are evaluated with ROUGE, a natural approach is to develop systems that maximize it.

Should novel automatic evaluation metrics be developed, our approach can still be applied, provided that the new metrics can be expressed as a function of the scores of individual sentences.

Structured Learning Compared to MDS approaches using structured learning, our problem-reduction has the important advantage that it considerably scales-up the available training data by working on sentences instead of documents/summaries pairs. Moreover, the task of sentence scoring is not dependent on arbitrary parameters such as the summary length which are inherently abstracted from the "summary worthiness" of individual textual units.

Error Propagation The first step of the framework is left to a ML component which can only produce approximate scores. Empirical results (in Figure 1 and Table 2) suggest that even with an

[8] The symbol * indicates that the difference compared to the previous best baseline is significant with $p \leq 0.05$.

imperfect first step, the subsequent optimization is able to produce high scoring summaries. However, it might be insightful to study rigorously and in greater detail the propagation of errors induced by the first step.

Other Metrics This work focused on maximizing ROUGE-N recall because it is a widely acknowledged automatic evaluation metric. ROUGE-N relies on reference summaries which forces us to perform an estimation step. In our framework, we use ML to estimate the individual scores of sentences without using reference summaries.

However, Louis and Nenkova (2013) proposed several alternative evaluation metrics for system summaries which do not need reference summaries. They are based on the properties of the system summary and the source documents alone, and correlate well with human evaluation. Some of them can even reach a correlation with human evaluation similar to the ROUGE-2 recall.

An example of such a metric is the Jensen-Shannon Divergence (JSD) which is a symmetric smoothed version of the Kullback-Leibler divergence. Maximizing JSD can not be solved exactly with an ILP because it can not be factorized into individual sentences. However, applying an efficient greedy algorithm or maximizing a factorizable relaxation might produce strong results as well (for example, a simple greedy maximization of Kullback-Leibler divergence already yields good results (Haghighi and Vanderwende, 2009)).

Future Work In this work, we developed a principled subset selection framework and empirically justified it. We focused on solving the second step of the framework while keeping the machine learning component as simple as possible. Essentially, our framework performs a modularization of the task of MDS, where all characteristics of the data and feature representations are pushed into a separate machine learning module – they should not affect the subsequent optimization step which remains fixed.

The promising results we obtained for summarization with a basic learner (see Section 4.3) encourage future work on plugging in more sophisticated supervised learners in our framework. For example, we plan to incorporate lexical-semantic information in the feature representation and leverage large-scale unsupervised pre-training. This direction is particularly promising because we have shown that we can expect significant performance gains for end-to-end MDS as the sentence scoring component improves.

6 Conclusion

We proposed a problem-reduction approach to extractive MDS, which performs a reduction to the problem of scoring individual sentences with their ROUGE scores based on supervised learning. We defined a principled discrete optimization problem for sentence selection which relies on an approximation of ROUGE. We empirically checked the validity of the approach on standard datasets and observed that even with a basic learner the framework produces promising results. The code for our optimizers is available at `github.com/UKPLab/acl2016-optimizing-rouge`.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group “Adaptive Preparation of Information from Heterogeneous Sources” (AIPHES) under grant No. GRK 1994/1.

References

Chinatsu Aone, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. 1995. A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 68–73. MIT Press, Cambridge, MA, USA.

Florian Boudin, Hugo Mougard, and Benot Favre. 2015. Concept-based Summarization using Integer Linear Programming: From Concept Pruning to Multiple Optimal Solutions. In Llus Mrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1914–1918, Lisbon, Portugal.

Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *Journal of Artificial Intelligence Research*, pages 457–479.

Dan Gillick and Benoit Favre. 2009. A Scalable Global Model for Summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, pages 10–18, Boulder, Colorado.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring Content Models for Multi-document Summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370.

Kai Hong and Ani Nenkova. 2014. 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 (EACL)*, pages 712–721, Gothenburg, Sweden.

Kai Hong, John Conroy, benoit Favre, Alex Kulesza, Hui Lin, and Ani Nenkova. 2014. A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1608–1616, Reykjavik, Iceland.

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive Summarization using Continuous Vector Space Models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 31–39, Gothenburg, Sweden.

Alex Kulesza and Ben Taskar. 2012. Determinantal Point Processes for Machine Learning. *Foundations and Trends in Machine Learning*, 5:123–286.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A Trainable Document Summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington, USA. Association for Computing Machinery.

Chen Li, Fei Liu, Fuliang Weng, and Yang Liu. 2013a. Document Summarization via Guided Sentence Compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 490–500, Seattle, Washington, USA.

Chen Li, Xian Qian, and Yang Liu. 2013b. Using Supervised Bigram-based ILP for Extractive Summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1004–1013, Sofia, Bulgaria.

Chen Li, Yang Liu, and Lin Zhao. 2015. Improving Update Summarization via Supervised ILP and Sentence Reranking. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1317–1322, Denver, Colorado.

Hui Lin and Jeff A. Bilmes. 2011. A Class of Submodular Functions for Document Summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 510–520, Portland, Oregon.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out at ACL*, pages 74–81, Barcelona, Spain.

Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. *Computational Linguistic*, 39(2):267–300, June.

Hans Peter Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research Development*, 2:159–165.

Inderjeet Mani and Eric Bloedorn. 1997. Multi-document Summarization by Graph Search and Matching. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, pages 622–628, Providence, Rhode Island. AAAI Press.

Ryan McDonald. 2007. A Study of Global Inference Algorithms in Multi-document Summarization. In *Proceedings of the 29th European Conference on IR Research*, pages 557–564, Rome, Italy. Springer-Verlag.

Rada Mihalcea. 2004. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In *Proceedings of the ACL 2004 on Interactive Poster and Demonstration Sessions*, ACLdemo '04, page 20, Barcelona, Spain.

Olof Mogren, Mikael Kågebäck, and Devdatt Dubhashi. 2015. Extractive Summarization by Aggregating Multiple Similarities. In *Recent Advances in Natural Language Processing*, pages 451–457, Hissar, Bulgaria.

George L. Nemhauser and Laurence A. Wolsey. 1978. Best Algorithms for Approximating the Maximum of a Submodular Set Function. *Mathematics of Operations Research*, 3(3):177–188.

George L. Nemhauser, Laurence A. Wolsey, and Marschall L. Fisher. 1978. An Analysis of Approximations for Maximizing Submodular Set FunctionsI. *Mathematical Programming*, 14:265–294.

Jun-Ping Ng, Praveen Bysani, Ziheng Lin, Min-Yen Kan, and Chew-Lim Tan. 2012. Exploiting Category-Specific Information for Multi-Document Summarization. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 2093–2108, Mumbai, India.

Hitoshi Nishikawa, Kazuho Arita, Katsumi Tanaka, Tsutomu Hirao, Toshiro Makino, and Yoshihiro Matsuo. 2014. Learning to Generate Coherent Summary with Discriminative Hidden Semi-Markov Model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1648–1659.

Paul Over, Hoa Dang, and Donna Harman. 2007. DUC in Context. *Information Processing and Management*, 43(6):1506–1520.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9, Montreal, Canada.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation, and User Studies. In *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, volume 4, pages 21–30, Seattle, Washington.

Ruben Sipos, Pannaga Shivaswamy, and Thorsten Joachims. 2012. Large-margin Learning of Submodular Summarization Models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 224–233, Avignon, France.

Hiroya Takamura and Manabu Okumura. 2010. Learning to Generate Summary as Structured Output. In *Proceedings of the 19th ACM international Conference on Information and Knowledge Management*, pages 1437–1440. Association for Computing Machinery.

Kristian Woodsend and Mirella Lapata. 2012. Multiple Aspect Summarization Using Integer Linear Programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, (EMNLP-CoNLL)*, pages 233–243, Jeju Island, Korea.

Wen-tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-document Summarization by Maximizing Informative Content-words. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, pages 1776–1782, Hyderabad, India. Morgan Kaufmann Publishers Inc.

Wenpeng Yin and Yulong Pei. 2015. Optimizing Sentence Modeling and Selection for Document Summarization. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 1383–1389, Buenos Aires, Argentina. AAAI Press.

A Supplemental Material

A.1 Recursive Expression of ROUGE-N

Let $S = \{s_i | i \leq m\}$ and $T = \{t_i | i \leq l\}$ be two sets of sentences, S^* the reference summary, and $\rho(X)$ denote the ROUGE-N score of the set of sentences X. Assuming that $\rho(S)$ and $\rho(T)$ are given, we prove the following recursive formula:

$$\rho(S \cup T) = \rho(S) + \rho(T) - \epsilon(S \cap T) \qquad (12)$$

For compactness, we use the following notation as well:

$$C_{X,S^*}(g) = min(F_X(g), F_{S^*}(g)) \qquad (13)$$

Proof: We have the following definitions:

$$\rho(S) = \frac{1}{R_N} \sum_{g \in S^*} \tilde{F}_{S,S^*}(g) \qquad (14)$$

$$\rho(T) = \frac{1}{R_N} \sum_{g \in S^*} \tilde{F}_{T,S^*}(g) \qquad (15)$$

$$\epsilon(S \cap T) = \frac{1}{R_N} \sum_{g \in S^*} max(C_{S,S^*}(g) + C_{T,S^*}(g) - F_{S^*}(g), 0) \qquad (16)$$

And by definition of ROUGE, the formula of $S \cup T$:

$$\rho(S \cup T) = \frac{1}{R_N} \sum_{g \in S^*} min(F_{S \cup T}(g), F_{S^*}(g)) \qquad (17)$$

In order to prove equation (12), we have to show that the following equation holds:

$$\sum_{g \in S^*} C_{S,S^*}(g) + \sum_{g \in S^*} C_{T,S^*}(g) - \sum_{g \in S^*} max(C_{S,S^*}(g) + C_{T,S^*}(g) - F_{S^*}(g), 0) = \sum_{g \in S^*} min(F_{S \cup T}(g), F_{S^*}(g)) \qquad (18)$$

It is sufficient to show:

$$\forall g \in S^*, C_{S,S^*}(g) + C_{T,S^*}(g) - max(C_{S,S^*}(g) + C_{T,S^*}(g) - F_{S^*}(g), 0) = min(F_{S \cup T}(g), F_{S^*}(g)) \qquad (19)$$

Let $g \in S^*$ be a n-gram. There are two possibilities:

- $F_S(g) + F_T(g) \leq F_{S^*}(g)$: g appears less times in $S \cup T$ than in the reference summary. It implies: $min(F_{S \cup T}(g), F_{S^*}(g)) = F_{S \cup T}(g) = F_S(g) + F_T(g)$. Moreover, all $F_X(g)$ are positive numbers by definition, and $F_S(g) \leq F_{S^*}(g)$ is equivalent to: $C_{S,S^*}(g) = min(F_S(g), F_{S^*}(g)) = F_S(g)$. Similarly, we have: $C_{T,S^*}(g) = min(F_T(g), F_{S^*}(g)) = F_T(g)$. Since $max(C_{S,S^*}(g) + C_{T,S^*}(g) - F_{S^*}(g), 0) = 0$, the equation (19) holds in this case.

- $F_S(g) + F_T(g) \geq F_{S^*}(g)$: g appears more frequently in $S \cup T$ than in the reference summary. It implies: $min(F_{S \cup T}(g), F_{S^*}(g)) = F_{S^*}(g)$. Here we have: $max(C_{S,S^*}(g) + C_{T,S^*}(g) - F_{S^*}(g), 0) = C_{S,S^*}(g) + C_{T,S^*}(g) - F_{S^*}(g)$, and it directly follows that equation (19) holds in this case as well.

Equation (19) has been proved, which proves (12) as well.

A.2 Expanded Expression of ROUGE-N

Let $S = \{s_i | i \leq m\}$ be a set of sentences and $\rho(S)$ its ROUGE-N score. We prove the following formula:

$$\rho(S) = \sum_{i=1}^{m} \rho(s_i) + \sum_{k=2}^{m} (-1)^{k+1} \Big(\sum_{1 \leq i_1 \leq \cdots \leq i_k \leq m} \epsilon^{(k)}(s_{i_1} \cap \cdots \cap s_{i_k}) \Big) \tag{20}$$

Proof: Let $g \in S^*$ be a n-gram in the reference summary, and $k \in [1, m]$ the number of sentences in which it appears. Specifically, $\exists \{s_{i_1}, \cdots, s_{i_k}\}, \forall s_{i_j} \in \{s_{i_1}, \ldots, s_{i_k}\}, g \in s_{i_j}$. In order to prove the formula (20), we have to find an expression for the $\epsilon^{(k)}$ that gives to g the correct contribution to the formula:

$$\frac{1}{R_N} min(F_S(g), F_{S^*}(g)) \tag{21}$$

First, we observe that g does not appear in the terms that contain the intersection of more than k sentences. Specifically, $\epsilon^{(t)}$ is not affected by g if $t \geq k$. However, g is affected by all the $\epsilon^{(t)}$ for which $t \leq k$.

Given that g appears in the sentences $\{s_{i_1}, \ldots, s_{i_j}\}$, we can determine the score attributed to g by the previous $\epsilon^{(t)}$ ($t \leq k$):

$$S^{(k-1)}(g) = \sum_{s \in \{s_{i_1}, \ldots, s_{i_k}\}} \rho(s) + \sum_{l=2}^{k} (-1)^{(l+1)} \sum_{1 \leq i_1 \leq \cdots \leq i_l \leq k} \epsilon^{(l)}(s_{i_1} \cap \cdots \cap s_{i_l})) \tag{22}$$

Now, g receives the correct contribution to the overall scores if $\epsilon^{(k)}$ is defined as follows:

$$\epsilon^{(k)}(s_{i_1} \cap \cdots \cap s_{i_j}) = \frac{1}{R} \sum_{g \in s_{i_1} \cap \cdots \cap s_{i_j}} min(C_{\{s_{i_1}, \ldots, s_{i_k}\}}(g), F_{S^*}(g)) - S^{(k-1)}(g) \tag{23}$$

Indeed, with this expression for $\epsilon^{(k)}$, the score of g is:

$$S^{(k-1)}(g) + \frac{1}{R_N} min(C_{\{s_{i_1}, \ldots, s_{i_k}\}}(g), F_{S^*}(g)) - S^{(k-1)}(g) \tag{24}$$

Which can be simplified to:

$$\frac{1}{R_N} min(C_{\{s_{i_1}, \ldots, s_{i_k}\}}(g), F_{S^*}(g)) \tag{25}$$

Since g appears only in the sentences $\{s_{i_1}, \ldots, s_{i_k}\}$, $\tilde{F}_{\{s_{i_1}, \ldots, s_{i_k}\}}(g) = F_S(g)$ and it follows that:

$$\frac{1}{R_N} min(C_{\{s_{i_1}, \ldots, s_{i_k}\}}(g), F_{S^*}(g)) = \frac{1}{R_N} min(F_S(g), F_{S^*}(g)) \tag{26}$$

This proves equation (20) because we observe that g will not be affected by any other terms. Every $\epsilon^{(t)}$ for $t \leq k$ including g is counted by $S^{(k-1)}$, and no other terms from $\epsilon^{(k)}$ will affect g because all the other terms $\epsilon^{(k)}$ should contain at least one sentence that is not in $\{s_{i_1}, \ldots, s_{i_k}\}$ and g would not belong to this intersection by definition.

Finally, it has been proved in the appendix A.1 that for $k = 2$, $\epsilon^{(2)}$ has a reduced form:

$$\epsilon^{(2)}(s_a \cap s_b) = \frac{1}{R_N} \sum_{g \in S^*} max(C_{s_a,S^*}(g) + C_{s_b,S^*}(g) - F_{S^*}(g), 0) \tag{27}$$

In the paper, we ignore the terms for $k \geq 2$, therefore we do not search for a reduced form for these terms.