Mining the Gaps: Towards Polynomial Summarization

Marina Litvak  
Sami Shamoon College of Engineering  
Beer Sheva, Israel  
marinal@sce.ac.il

Natalia Vanetik  
Sami Shamoon College of Engineering  
Beer Sheva, Israel  
natalyav@sce.ac.il

Abstract

The problem of text summarization for a collection of documents is defined as the problem of selecting a small subset of sentences so that the contents and meaning of the original document set are preserved in the best possible way. In this paper we present a linear model for the problem of text summarization, where a summary preserves the information coverage as much as possible in comparison to the original document set. We reduce the problem of finding the best summary to the problem of finding the point on a convex polytope closest to the given hyperplane, and solve it efficiently with the help of fractional (polynomial-time) linear programming. The experimental results show the superiority of our approach over most of the systems participating in the generic multi-document summarization task (MultiLing) of the TAC 2011 competition.

1 Introduction

Automated text summarization is an active field of research in various communities like Information Retrieval (IR), Natural Language Processing (NLP), and Text Mining (TM).

Some authors reduce summarization to the maximum coverage problem (Takamura and Okumura, 2009; Gillick and Favre, 2009) that, despite a great performance, is known as NP-hard (Khuller et al., 1999). Linear Programming helps to find an accurate approximated solution to this problem and became very popular in summarization field in the last years (Gillick and Favre, 2009; Woodsend and Lapata, 2010; Hitoshi Nishikawa and Kikui, 2010; Makino et al., 2011). However, most mentioned works use exponential number of constrains or Integer Linear Programming which is an NP-hard problem.

Trying to solve a trade-off between summary quality and time complexity, we propose a novel summarization model solving the approximated maximum coverage problem by linear programming in polynomial time. We measure information coverage by terms and strive to obtain a summary that preserves the optimal value of the chosen objective function as much as possible in comparison to the original document. Various objective functions combining different parameters like term’s position and its frequency are introduced and evaluated.

Our method ranks and extracts significant sentences into a summary and it can be generalized for both single-document and multi-document summarization. Also, it can be easily adapted to cross-lingual/multilingual summarization.

Formally speaking, in this paper we introduce (1) a novel text representation model expanding a classic Vector Space Model (Salton et al., 1975) to Hyperplane and Half-spaces, (2) re-formulated extractive summarization problem as an optimization task and (3) its solution using linear or quadratic programming. The main challenge of this paper is a new text representation model making possible to represent an exponential number of extracts without computing them explicitly, and finding the optimal one by simple minimizing a distance function in polynomial time.

2 Our Method

2.1 Definitions

We are given a set of sentences $S_1, ..., S_n$ derived from a document or a cluster of related documents. Meaningful words in these sentences are entirely described by terms $T_1, ..., T_m$. Our goal is to find a
subset \( S_i, \ldots, S_k \) consisting of sentences such that (1) there are at most \( N \) terms in these sentences, (2) term frequency is preserved as much as possible w.r.t. the original sentence set, (3) redundant information among \( k \) selected sentences is minimized.

We use the standard sentence-term matrix, \( A = (a_{ij}) \) of size \( m \times n \), for initial data representation, where \( a_{ij} = k \) if term \( T_i \) appears in the sentence \( S_j \) precisely \( k \) times. Here, columns of \( A \) describe sentences and rows describe terms. Since we are not interested in redundant sentences, in the case of multi-document summarization, we can initially select meaningful sentences by clustering all the columns as vectors in \( \mathbb{R}^n \) and choose a single representative from each cluster. In this case columns of \( A \) describe representatives of sentence clusters. The total number of words (term appearances) in the document, denoted by \( S \), can be computed from the matrix \( A \) as

\[
S = \sum_i \sum_j a_{ij} \tag{1}
\]

**Example 1.** Given the following text of \( n = 3 \) sentences and \( m = 5 \) (normalized) terms:

- \( S_1 = \) A fat cat is a cat that eats fat meat.
- \( S_2 = \) My cat eats fish but he is a fat cat.
- \( S_3 = \) All fat cats eat fish and meat.

Matrix \( A \) corresponding to the text above has the following shape:

\[
T_1 = \text{“fat”} \begin{bmatrix} a_{11} = 2 & a_{12} = 1 & a_{13} = 1 \end{bmatrix} \\
T_2 = \text{“cat”} \begin{bmatrix} a_{21} = 2 & a_{22} = 2 & a_{23} = 1 \end{bmatrix} \\
T_3 = \text{“eat”} \begin{bmatrix} a_{31} = 1 & a_{32} = 1 & a_{33} = 1 \end{bmatrix} \\
T_4 = \text{“fish”} \begin{bmatrix} a_{41} = 0 & a_{42} = 1 & a_{43} = 1 \end{bmatrix} \\
T_5 = \text{“meat”} \begin{bmatrix} a_{51} = 1 & a_{52} = 0 & a_{53} = 1 \end{bmatrix}
\]

where \( a_{ij} \) are term counts. The total count of terms in this matrix is

\[
S = \sum_{i=1}^{5} \sum_{j=1}^{3} a_{ij} = 16
\]

Our goal is to find subset \( t_1, \ldots, t_k \) of \( A \)'s columns so that the chosen submatrix represents the best possible summary under some constraints. Since it is hard to determine what is the best summary mathematically (this task is usually left to human experts), we wish to express summary quality as a linear function of the underlying matrix. We strive to find a summary that gives an optimal value once the function in question has been determined.

### 2.2 Text Preprocessing

In order to build the matrix and then the polytope model, one needs to perform the basic text preprocessing including sentence splitting and tokenization. Also, additional steps like stopwords removal, stemming, synonym resolution, etc. may be performed for resource-rich languages. Since the main purpose of these methods is to reduce the matrix dimensionality, the resulted model will be more efficient.

### 2.3 Polytope as a document representation

We represent every sentence by a hyperplane, and all sentences derived from a document form a hyperplane intersections (polytope). Then, all possible extracts can be represented by subplanes of our hyperplane intersections and as such that are not located far from the boundary of the polytope. Intuitively, the boundary of the resulting polytope is a good approximation for extracts that can be generated from the given document. We view every column of the sentence-term matrix as a linear constraint representing a hyperplane in \( \mathbb{R}^{mn} \). An occurrence of term \( t_i \) in sentence \( S_j \) is represented by variable \( x_{ij} \). The maximality constraint on the number of terms in the summary can be easily expressed as a constraint on the sum of these variables.

**Example 2.** This example demonstrates variables corresponding to the \( 5 \times 3 \) matrix \( A \) of Example 1.

<table>
<thead>
<tr>
<th></th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( T_1 )</td>
<td>( x_{11} )</td>
<td>( x_{12} )</td>
<td>( x_{13} )</td>
</tr>
<tr>
<td>( T_2 )</td>
<td>( x_{21} )</td>
<td>( x_{22} )</td>
<td>( x_{23} )</td>
</tr>
<tr>
<td>( T_3 )</td>
<td>( x_{31} )</td>
<td>( x_{32} )</td>
<td>( x_{33} )</td>
</tr>
<tr>
<td>( T_4 )</td>
<td>( x_{41} )</td>
<td>( x_{42} )</td>
<td>( x_{43} )</td>
</tr>
<tr>
<td>( T_5 )</td>
<td>( x_{51} )</td>
<td>( x_{52} )</td>
<td>( x_{53} )</td>
</tr>
</tbody>
</table>

Every sentence in our document is a hyperplane in \( \mathbb{R}^{mn} \), defined with columns of \( A \) and variables representing terms in sentences:

\[
A[:, j] = [a_{i1j}, \ldots, a_{imj}]
\]

\[
x_j = [x_{1,j}, \ldots, x_{mn,j}] \text{ for all } 1 \leq j \leq n
\]

We define a system of linear inequalities

\[
A[:, j] \cdot x_j^T = \sum_{i=1}^{m} a_{ij} x_{ij} \leq A[:, j] \cdot 1^T = \sum_{i=1}^{m} a_{ij}
\]

Every inequality of this form defines a hyperplane \( H_j \) and it lower half-space specified by equation (2):

\[
A[:, j] \cdot x_j^T = A[:, j] \cdot 1^T
\]
and with normal vector \( n = (n_{xy}) \)

\[
n_{xy} = \begin{cases} 
  a_{xy} & 1 \leq x \leq m \land y = j \\
  0 & \text{otherwise.}
\end{cases}
\] (3)

To say that every term is either present or absent from the chosen extract, we add constraints \( 0 \leq x_{ij} \leq 1 \). Intuitively, entire hyperplane \( H_i \) and therefore every point \( p \in H_i \) represents sentence \( S_i \). Then a subset of \( r \) sentences is represented by intersection of \( r \) hyperplanes.

**Example 3.** Sentence-term matrix \( A \) of Example 1 defines the following hyperplane equations.

\[
\begin{align*}
H_1 : & \quad 2x_{11} + 2x_{21} + x_{31} + x_{51} = 2 + 2 + 1 + 1 = 6 \\
H_2 : & \quad x_{12} + 2x_{22} + x_{32} + x_{42} = 5 \\
H_3 : & \quad x_{13} + x_{23} + x_{33} + x_{43} + x_{53} = 5
\end{align*}
\]

Here, a summary consisting of the first and the second sentence is expressed by the intersection of hyperplanes \( H_1 \) and \( H_2 \). Figure 1 shows how a two-dimensional projection of hyperplanes \( H_1, H_2, H_3 \) and their intersections look like.

### 2.4 Summary constraints

We express summarization constraints in the form of linear inequalities in \( \mathbb{R}^{mn} \), using the columns of the sentence-term matrix \( A \) as linear constraints. Maximal constraint on the number of terms in the summary can be easily expressed as a constraint on the sum of term variables \( x_{ij} \).

\[
\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} \leq T_{\text{max}}
\] (4)

**Example 4.** Equation (4) for Example 1, \( T_{\text{max}} = 11 \) has the form

\[
\begin{align*}
0 \leq x_{ij} & \leq 1, \forall i, j \\
\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} & \leq 11
\end{align*}
\]

Additionally, we may have constraints on the maximal \( W_{\text{max}} \) number of words in the summary. We take into account only words that remain in the text after stop-word removal and stemming. The difference between the number terms and the number of words in a summary is that a single term can appear more than once in a sentence. Therefore, the total number of words in the text is expressed by summing up the elements of its term-count matrix. Therefore, maximality constraints for words are expressed by the following linear inequality.

\[
\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} x_{ij} \leq W_{\text{max}}
\] (5)

**Example 5.** Equation (5) for the sentence-term matrix of Example 1 for \( W_{\text{max}} = 11 \) has the form

\[
\begin{align*}
2x_{11} + 2x_{21} + x_{31} + x_{51} + & \\
+ x_{12} + x_{22} + 2x_{32} + x_{42} + & \\
+ x_{13} + x_{23} + x_{33} + x_{43} + x_{53} & \leq 11
\end{align*}
\]

### 2.5 The polytope model

Having defined linear inequalities that describe each sentence in a document separately and the total number of terms in sentence subset, we can now look at them together as a system:

\[
\begin{align*}
\sum_{i=1}^{m} a_{i1} x_{i1} & \leq \sum_{i=1}^{m} a_{i1} \\
\vdots & \\
\sum_{i=1}^{m} a_{in} x_{in} & \leq \sum_{i=1}^{m} a_{in} \\
\sum_{j=1}^{n} x_{ij} & \leq T_{\text{max}} \\
\sum_{j=1}^{n} a_{ij} x_{ij} & \leq W_{\text{max}} \\
0 & \leq x_{ij} \leq 1
\end{align*}
\] (6)

First \( n \) inequalities describe sentences \( S_1, \ldots, S_n \), the next two inequalities describes constraints on the total number of terms and words in a summary, and the final constraint determines upper and lower boundaries for all sentence-term variables. Since every inequality in the system (6) is linear, the entire system describes a convex polyhedron in \( \mathbb{R}^{mn} \), which we denote by \( P \). Faces of \( P \) are determined by intersections of hyperplanes defined in (6).

### 2.6 Objectives and summary extraction

We assume here that the surface of the polyhedron \( P \) is a suitable representation of all the possible sentence subsets (its size, of course, is not polynomial in \( m \) and \( n \) since the number of vertices of \( P \) can reach \( O(2^n) \)). Fortunately, we do not need to scan the whole set of \( P \)'s surfaces but rather to find
the point on P that optimizes the chosen objective function. Table 1 contains four different objective functions\(^3\) that we used for summarization, along with descriptions of the changes in the model that were required for each function.

Since the LP method not only finds the minimal distance but also presents an evidence to that minimality in the form of a point \(x = (x_{ij})\), we use the point’s data to find what sentences belong to the chosen summary. We check which equations of \(H_i\) the point \(x\) satisfies as equalities. If an equality holds, \(x\) lies on \(H_i\) and therefore the sentence \(S_i\) is contained in the summary. This test is straightforward and takes \(O(mn)\) time. In a case of insufficient summary length, the sentences nearest to the point \(x\) are extracted to a summary in a greedy manner.

<table>
<thead>
<tr>
<th>Function</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
</table>
| Maximal Weighted Term Sum (OBJ\(_1\)) | \(\max \sum_{i=1}^{m} w_i t_i\) \(t_i = \sum_{j=1}^{n} x_{ij}\) | Maximizes the information coverage as a weighted term sum. We used the following types of term weights \(w_i\):
1. \(\text{POS}_\text{EQ}\), where \(w_i = 1\) for all \(i\);
2. \(\text{POS}_\text{F}\), where \(w_i = \frac{1}{\text{app}(i)}\) and \(\text{app}(i)\) is the index of a sentence in the document where the term \(T_i\) first appeared;
3. \(\text{POS}_\text{B}\), where \(w_i = \max\left\{\frac{1}{\text{app}(i)} \cdot n_{\text{app}(i)} / \sum_{j=1}^{n} n_{\text{app}(j)}\right\}\);
4. \(\text{TF}\), where \(w_i = \text{tf}(i)\) and \(\text{tf}(i)\) is the term frequency of term \(T_i\);
5. \(\text{TFISF}\), where \(w_i = \text{tf}(i) \cdot \text{isf}(i)\) and \(\text{isf}(i)\) is the inverse sentence frequency of \(T_i\). |
| Distance Function (OBJ\(_2\)) | \(\min \sum_{i=1}^{m} (t_i - p_i)^2\) \(t_i = \sum_{j=1}^{n} x_{ij}\) and \(p_i = 1\), or \(t_i = \frac{\sum_{j=1}^{n} t_j}{n}\) and \(p_i = \text{tf}(i)\) | Minimizes the Euclidean distance between terms \(t = (t_1, \ldots, t_m)\) (a point on the polytope \(P\) representing a generated summary) and the vector \(p = (p_1, \ldots, p_m)\) (expressing document properties we wish to preserve and representing the “ideal” summary). We used the following options for \(t\) and \(p\) representation:
1. \(\text{MTC}\), where \(t\) is a summary term count vector and \(p\) contains all the terms precisely once, thus minimizing repetition but increasing terms coverage.
2. \(\text{MTF}\), where \(t\) contains term frequencies of terms in a summary and \(p\) contains term frequency for terms in documents. |
| Sentence Overlap (OBJ\(_3\)) | \(\min \sum_{i=1}^{m} \sum_{j=1}^{n} ovl_{jk}\) \(ovl_{jk} = \frac{|S_j \cap S_k|}{\min(|S_j|, |S_k|)}\) | Minimizes the Jaccard similarity between sentences in a summary (denoted by \(ovl_{jk}\) for \(S_j\) and \(S_k\)). \(w(a_{ij}, a_{ik})\) is 1 if the term \(T_i\) is present in both sentences \(S_j\) and \(S_k\) and is 0 otherwise. |
| Maximal Bigram Sum (OBJ\(_4\)) | \(\max \sum_{i,j} b_{ij}\) where \(0 \leq b_{ij} \leq 1\) | Maximizes the information coverage as a bigram sum. Variable \(b_{ij}\) is defined for every bigram \((T_i, T_j)\) in the text. |

We conducted the experiments on the MultiLing 2011 (Giannakopoulos et al., 2011) English dataset. MultiLing dataset consists of 10 document sets, 10 documents each one, in seven languages. The original news articles in English were taken from WikiNews\(^5\), organized into 10 sets, and then summarized. According to the MultiLing summarization task, all systems must generate summaries in size of 250 words at most. Eight systems (ID1-ID8) participated in the pilot and compared to the global baseline (ID9) and the global topline (ID10) systems. Systems A,B and C de-\(^4\)The software is available upon request.

\(^3\)Since our approach is unsupervised, there is no possibility and meaning to use ROUGE, that needs Gold Standard, as an objective.

\(^5\)http://en.wikinews.org/wiki/
note summaries manually created by human ex-
erts. The choice of this dataset is argued by future plans to adapt and evaluate the introduced system to multiple languages.

The automatic summarization evaluation package, ROUGE (Lin, 2004), is used to evaluate the effectiveness of our approach vs. 10 summarizers participated in the MultiLing pilot of the TAC 2011 competition. For fair comparison, only first 250 words\(^6\) were considered in ROUGE statistics. The recall scores of ROUGE-N for \(N \in \{1, 2, 3, 4\}\), ROUGE-W-1.2, and ROUGE-SU4 which are based on \(N\)-gram, Weighted Longest Common Subsequence (WLCS), and Skip-bigram plus unigram, with maximum skip-distance of 4, matching between system summaries and reference summaries, respectively, are reported in Table 2 below.

### 3.1 Experimental Results

As it can be seen from Table 2, our model using unweighted term sum (\(OBJ_1^{POS,EQ}\)) as an objective function outperforms most of the systems – 6 systems in terms of ROUGE-1, ROUGE-2, ROUGE-SU4 and ROUGE-W-1.2, and 8 systems in terms of ROUGE-3 and ROUGE-4. Conversely to our expectations, adding any type of weights to \(OBJ_1\) reduces its performance. Minimizing repetition while increasing terms coverage (\(OBJ_2^{MTC}\)) shares the same rank with \(OBJ_1^{POS,EQ}\) for most ROUGE metrics. Minimizing distance to a document term frequency vector (\(OBJ_2^{MTF}\)) performs worse – it outperforms 3, 4, 5, 6, 5 and 3 systems in terms of ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-SU4 and ROUGE-W-1.2, respectively. Sentence overlap (\(OBJ_3\)) and maximal bigram sum (\(OBJ_4\)) have very close scores, outperforming 3, 5, 6, 5 and 3 systems in terms of ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-SU4 and ROUGE-W-1.2, respectively. Generally, optimizing the most of introduced functions generates the near-quality summaries. All functions perform better then the baseline (ID9) system.\(^7\)

### 4 Conclusions and Future Work

In this paper we present a linear programming model for the problem of extractive summariza-

---

\(^6\)ROUGE.pl -a -x -2 -4 -u -c 95 -e data -r 1000 -n 4 -f A -p 0.5 -t 0 -d -l 250

\(^7\)We did not perform tests of statistical significance due to too many comparisons (10 systems vs. 10 objective functions), leaving it as a future work.


