Enhancing Diversity, Coverage and Balance for Summarization through Structure Learning

Liangda Li[†], Ke Zhou[†], Gui-Rong Xue[†*], Hongyuan Zha[‡], and Yong Yu[†]
[†]Dept. of Computer Science and Engineering
Shanghai Jiao-Tong University
No. 800, Dongchuan Road, Shanghai, China 200240
{ldli ,zhouke, grxue, yyu}@apex.sjtu.edu.cn

[‡]College of Computing
Georgia Institute of Technology
Atlanta, GA 30032
zha@cc.gatech.edu

ABSTRACT

Document summarization plays an increasingly important role with the exponential growth of documents on the Web. Many supervised and unsupervised approaches have been proposed to generate summaries from documents. However, these approaches seldom simultaneously consider summary diversity, coverage, and balance issues which to a large extent determine the quality of summaries. In this paper, we consider extract-based summarization emphasizing the following three requirements: 1) diversity in summarization, which seeks to reduce redundancy among sentences in the summary; 2) sufficient coverage, which focuses on avoiding the loss of the document's main information when generating the summary; and 3) balance, which demands that different aspects of the document need to have about the same relative importance in the summary. We formulate the extract-based summarization problem as learning a mapping from a set of sentences of a given document to a subset of the sentences that satisfies the above three requirements. The mapping is learned by incorporating several constraints in a structure learning framework, and we explore the graph structure of the output variables and employ structural SVM for solving the resulted optimization problem. Experiments on the DUC2001 data sets demonstrate significant performance improvements in terms of F_1 and ROUGE metrics.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-abstracting methods

General Terms

Algorithm, Experimentation, Performance

Keywords

summarization, diversity, coverage, balance, structural SVM

*Gui-Rong Xue is the corresponding author.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

1. INTRODUCTION

Document summarization is the process of generating a short version of a given document to indicate its main topics. With the rapid growth of the Web, summarization has proved to be an essential task in many text processing areas. For example, in response to a user query, search engines often provide short summaries in the form of snippets for each document in the result list. In this way, users can save time by browsing the summaries before deciding whether or not to read the whole documents.

Document summarization can generally be categorized as abstract-based and extract-based [26]. Abstract-based summarization can be seen as a reproduction of the original document in a new way, while extract-based summarization focuses on extracting sentences from the original document [13, 16, 17, 26]. In this paper, we consider generic extract-based summarization of a single document.

Several learning-based methods have been proposed for extract-based summarization. They usually utilize a set of features constructed from the document and extract a subset of sentences from the document using some machine learning methods. For example, some approaches consider extract-based summarization as a binary classification problem, and use the information contained in each sentence and its popularity to decide whether it should be included in the summary. Unfortunately, those approaches tend to give rise to serious problems such as redundancy, unbalance and low recall in the generated summaries.

According to our observations, we argue that an effective summarization method should properly consider the following three key requirements:

- **Diversity:** A good document summary should be concise and contain as few redundant sentences as possible, i.e., two sentences providing similar information should not be both present in the summary. In practice, enforcing diversity in summarization can effectively reduce redundancy among the sentences.
- **Coverage:** The summary should contain every important aspects of the document. By taking coverage into consideration, the information loss in summarization can be minimized.

- **Balance:** The summary should emphasize the various aspects of the document in a balanced way. An unbalanced summary usually leads to serious misunderstanding of the general idea of the original document.

The goal of this paper is to systematically explore the above three aspects through a structure learning framework. The main idea is to utilize structural SVM with three types of constraints to enforce diversity, coverage and balance in the generated summaries. Specifically, based on the relations between sentences, we propose a so-called independence graph to model the structure of a given document, aiming at representing the dissimilarity between pairs of sentences and reducing the search space in the structure learning process. We also adapt the cutting plane algorithm to solve the resulting optimization problem and then use the trained model for summary generation.

A set of experiments is conducted on the DUC2001 data sets to evaluate our proposed method. Firstly, we compare the performance of our method with several state-of-the-art supervised and unsupervised methodes for summarization. Secondly, we evaluate the effectiveness of our three proposed constraints on enforcing diversity, coverage and balance of summaries. The experimental results show that explicit enforcement of diversity, coverage and balance results in significant improvements over state-of-the-art summarization methods. The two main contributions of our work are:

1. We take the issues of diversity, coverage and balance into consideration in extract-based summarization and enforce them through the construction of three types of constraints.

2. We propose an efficient structure learning framework to solve the summarization task, utilizing structural SVM with effective modeling of sentence relationships using independence graphs.

The rest of the paper is organized as follows: Section 2 presents related works. We cast extract-based summarization as a structure learning problem in Section 3. In Section 4, we discuss in detail the requirements of a high-quality summary, emphasizing diversity, coverage and balance issues, and how to employ constraints to handle them. Section 5 presents the details of our structure learning method including the construction of the independence graphs. A presentation of several useful features for the training process is also discussed. In section 6, we focus on the experimental results together with some analysis. We conclude the paper with some pointers to future research directions in Section 7.

2. RELATED WORKS

Document summarization has been an active topic in many research areas, such as natural language processing, information retrieval and machine learning. We discuss prior works and put our contributions in context.

Abstract-based summarization is an important problem in natural language processing. Knight and Marcu proposed a framework to generate summaries by creating sentences from the document's original content [16]. In addition, Jing and McKeown developed a cut-and-paste-based text summarizer, which modifies extracted sentences by discarding unimportant words and produces new sentences by merging result phrases [13]. Most of those methods focus on producing sentences from extracted sentences.

Extract-based summarization is usually viewed as a machine learning problem: selecting a subset of sentences from a given document. Several supervised learning methods have been developed for training accurate models for extract-based summarization. For example, the SVM-based method focuses on constructing a decision boundary between summary sentences and non-summary sentences [24]. However, it relies on the assumption that sentences are independent from each other and ignores the correlation between sentences. The HMM-based method proposed in [5] relaxes this assumption by modeling the relations between different sentences through hidden Markov models. Unfortunately, the training process of the HMM-based method becomes intractable when the feature space becomes very large. The CRF-based method, proposed by Shen et al. [26], gives a better solution to the problem of sentence dependency, leading to a sound result compared with other approaches.

Many unsupervised methods also contribute greatly to document summarization. Zha in [32] proposed a mutual reinforcement principle for sentence extraction using the idea of HITS [15]. Several related methods such as TextRank [21], LexPageRank [7] and CollabSum [30] gain remarkable performance also by exploring methodologies used in PageRank [2] and HITS [15]. These methods mostly focus on the dependence, for instance, similarity, between sentences of the same document or within multiple documents. Several clustering methods have also been employed for preprocessing, which indeed improve performance.

Several methods also cast the extract-based summarization as a sentence ranking problem, those include ranking through standard IR methods or identifying semantically important sentences by employing the latent semantic analysis technique [10]. Several learning to rank methods such as ranking SVM, support vector regression and gradient boosted decision trees are also applied to the summarization task [19].

Nomoto and Matsumoto take the diversity issue into consideration with a preprocessing step in which the sentences of a given document are first clustered into several groups [23]. CollabSum [30] reduces sentence redundancy by discarding the highly overlapping sentences with already extracted highly ranked sentences. Goldstein et al. consider this redundancy issue by employing a metric for reducing redundancy and maximizing diversity in the selected passages [9]. An earlier work by Carbonell and Goldstein [3] uses the idea of maximum marginal relevance for document reordering. In fact, most supervised methods which claim to deal with the diversity issue follows similar approaches by using a pre-defined fixed criterion. Our approach distinguishes itself by directly incorporating the diversity requirement into the training process.

The coverage issue is also important in summarization. IBM's many aspects document summarization tool takes this issue into consideration by automatically highlighting a set of sentences that are expected to cover the different aspects of the document's content [1]. However, to the best of our knowledge, most state-of-art approaches [17, 19, 26] employ 0-1 loss functions to measure the coverage of the ground-truth summary sentences. This kind of measurement is quite coarse and it does not capture a significant part of

the information available at the word level. Our method differs from this by introducing a systematic measurement to make full use of the information contained in training data to enforce the coverage ratio of the summarization.

We also mention that summarization has always had a wide range of applications for web documents [27]. Compared with generic text summarization, web document summarization usually explore more data sources, such as clickthrough data, metadata, or hyperlinks. Summarization is also used for web document classification [25], and particular for providing the snippets for search results.

The Diversity issue has be researched in many areas other than summarization. From the point of view of searching for diversified results, Yue and Joachims [31] have employed structural SVM to predict diverse subsets using loss functions to penalize low diversity. The approach taken by the DD-PREF modeling language tackles the diversity problem by measuring diversity on a feature basis [6, 29]. As another example, a framework presented by Clarke et al. rewards diversity in retrieval evaluation [4]. Besides, Kennedy and Naaman also propose a tool to generate diverse image search results [22]. Our method differs from above methods by incorporating diversity as a constraint for the training process, which proves to be a novel solution for the summarization task.

3. PROBLEM FORMULATION

Given a document $\mathbf{x} = \{x_1, x_2, \ldots, x_n\} \in \mathcal{X}$, where x_i represents the i-th sentence in the document, and $\mathcal{X}$ is the space of all documents, the summarization task is to predict a subset $\mathbf{y}$ from the space of all possible subsets $\mathcal{Y}$.

The general approach we pursue is supervised in nature, learning a summarization model from a set of training examples. To this end, assume that we have a set of labeled training data,

$$\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \mid i = 1, \cdots, n\},$$

where $\mathbf{y}^{(i)}$ is the ground-truth summary of the document $\mathbf{x}^{(i)}$. Given the training set, our goal is to construct a discriminant function

$$\mathcal{F}(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R},$$

which judges whether the subset $\mathbf{y}$ is a suitable summary for the document $\mathbf{x}$. The higher $\mathcal{F}(\mathbf{x}, \mathbf{y})$ is, the better $\mathbf{y}$ summarize the document $\mathbf{x}$. Therefore, we can predict the summary of a document $\mathbf{x}$ by maximizing $\mathcal{F}(\mathbf{x}, \mathbf{y})$ over the subset $\mathbf{y} \in \mathcal{Y}$. Formally,

$$\mathbf{y}^* = \underset{\mathbf{y} \in \mathcal{Y}}{\operatorname{argmax}} \mathcal{F}(\mathbf{x}, \mathbf{y}). \tag{1}$$

We describe each pair $(\mathbf{x}, \mathbf{y})$ through a feature vector $\Psi(\mathbf{x}, \mathbf{y})$ the exact form of which will be discussed later. The discriminant function $\mathcal{F}(\mathbf{x}, \mathbf{y})$ is assumed to be linear in the feature vector $\Psi(\mathbf{x}, \mathbf{y})$, i.e.,

$$\mathcal{F}(\mathbf{x}, \mathbf{y}) = \mathbf{w}^T \Psi(\mathbf{x}, \mathbf{y}) \tag{2}$$

To measure the summarization performance, a loss function

$$\Delta(\mathbf{y}, \bar{\mathbf{y}}) : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathcal{R}$$

is employed to quantify the penalty of the predicted summary $\bar{\mathbf{y}}$ compared with the ground-truth summary $\mathbf{y}$. In our study, a loss function similar to F_1 measure is applied:

$$\Delta(\mathbf{y}, \bar{\mathbf{y}}) = \frac{2pr}{p+r}, \quad p = \frac{< \mathbf{y}, \bar{\mathbf{y}} >}{< \bar{\mathbf{y}}, \bar{\mathbf{y}} >}, \quad r = \frac{< \mathbf{y}, \bar{\mathbf{y}} >}{< \mathbf{y}, \mathbf{y} >}. \tag{3}$$

Here, given two subsets $\mathbf{a}$ and $\mathbf{b}$, $< \mathbf{a}, \mathbf{b} >$ denotes the number of common elements they share.

4. STRUCTURE LEANING FRAMEWORK

In this study, we will explore structural SVM to train a robust model for the summarization task. For a given training set $\{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \mid i = 1, \cdots, n\}$, structural SVM is employed to learn a weight vector $\mathbf{w}$ for the discriminant function $\mathcal{F}(\mathbf{x}, \mathbf{y})$ through the following quadratic programing problem [8, 28, 31]:

Optimization Problem 1. (Structural SVM)

$$\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i, \tag{4}$$

subjected to:

$$\forall i, \forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}^{(i)}, \xi_i \geq 0,$$
$$\mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T \Psi(\mathbf{x}^{(i)}, \mathbf{y}) + \Delta(\mathbf{y}^{(i)}, \mathbf{y}) - \xi_i.$$

In Equation (4), the parameter c controls the tradeoff between the model complexity $\frac{1}{2}\|\mathbf{w}\|^2$ and $\sum_{i=1}^{n} \xi_i$, the sum of the slack variables ξ_i. The constraints for the optimization problem enforce the fact that the ground-truth summary $\mathbf{y}^{(i)}$ should have a higher value of discriminant function than other alternatives $\mathbf{y} \in \mathcal{Y}$.

As discussed above, we focus on training a summarization model which can enforce the diversity, coverage and balance of a summary. This can be achieved by introducing three additional types of constraints to the optimization problem defined in Equation (4). In the following section, we first define the notion of subtopics of a document and then describe our constraints for diversity, coverage and balance in terms of subtopics.

4.1 Constraints Based on Subtopic Set

Documents tend to contain several subtopics, and each subtopic can be represented by a cluster of sentences [11, 12]. Our constraints are based on this notion of subtopics. One important thing to notice is that the structure of subtopics of documents are only used in the training process through additional constraints in the optimization problem (4), it is not used in the testing phase when we need to generate a summary for a new document.

4.1.1 The Subtopic Set

Each subtopic of a document generally consists of a subset of the sentences of the document. For each document, we also define its subtopic set T as a set of subtopics that covers the document, i.e., $T = \{t_1, t_2, \ldots, t_k\}$, where subtopic t_i is associated with set of words. We now define cover(t_i, s) as the degree of coverage of the sentence s for the subtopic t_i:

$$\text{cover}(t_i, s) = \frac{|t_i \cap s|}{|t_i|}. \tag{5}$$

Here s is an arbitrary sentence in the document. Specifically, cover(t_i, s) represents the proportion of the words in the subtopic t_i that are also present in the sentence s. Furthermore, for each sentence s, its coverage of the give subtopic

set T can be measured by a vector $\mathbf{v} = (v_1, \ldots, v_k)$, where $v_i = \text{cover}(t_i, s)$.

It seems that we need to carry out a preprocessing step to construct the subtopic set for each document in the training set. However, this is not necessary in our context, because for each training document $\mathbf{x}^{(j)}$, we have the ground-thruth summary $\mathbf{y}^{(j)}$. Therefore, we simply use the i-th sentence in the ground-truth summary as the subtopic t_i for the document.

4.1.2 Constraints for Enforcing Diversity

Diversity argues a summary should not contain similar sentences. In other words, sentences in a summary should have little overlap with one another in order to reduce redundancy. Formally, we enforce diversity with the following constraint:

Constraint 1.

$$\mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \sum_{y \in \mathbf{y}^{(i)}} \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, y) - \mu + \xi_i \quad (6)$$

Generally, adding the slack variable μ tends to give slightly better performance.

In Equation (6), for the sentences in the ground-truth summary, the sum of their unique score $\mathbf{w}^T\Psi(\mathbf{x}, y)$ should be no more than the overall score when they are regarded as a whole set. In other words, we prefer summaries with each sentence focusing on different subtopics. In this way, the commonly shared features will be associated with relatively low weight. As the sentences similar to each other usually share lots of those features, a sentence set with less redundancy tends to be predicted.

4.1.3 Constraints for Enforcing Coverage

Coverage means that the generated summary should cover all subtopics as much as possible. Poor subtopic coverage is usually manifested by absence of some summary sentences. The following constraint is employed to enforce coverage:

Constraint 2.

$$\mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}) + (1 - \| \sum_{y_i \in \mathbf{y}} \mathbf{v}_i \|) - \xi_i. \quad (7)$$

As defined above, $\mathbf{v}_i$ denotes the coverage of y_i, the i-th sentence in summary $\mathbf{y}$, of the subtopics of a given document.[1]

For a given summary $\mathbf{y}$, $(1 - \| \sum_{y_i \in \mathbf{y}} \mathbf{v}_i \|)$ quantifies loss of coverage of the subtopics. As in (7), a large amount of loss of coverage leads to a relatively low score of $\mathbf{w}^T\Psi(\mathbf{x}, \mathbf{y})$. When predicting, sentences covering more subtopics tend to be extracted.

4.1.4 Constraints for Enforcing Balance

Balance requires that the generated summary should have relatively equal degree of coverage for each subtopic. The extraction of unbalanced information tend to cause heavy loss of information, given that the size of the summary is limited. We address this using the following constraint:

Constraint 3.

$$\mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}) + \sum_{j=1}^{k} (u_j - \bar{u})^2 - \xi_i, \quad (8)$$

[1]For a vector $\mathbf{u}$, $\|\mathbf{u}\|$ denotes the 2-norm of $\mathbf{u}$.

where k is the number of subtopics, $\mathbf{u} = \sum_{y_i \in \mathbf{y}} \mathbf{v}_i$, u_j is the j-th component of the vector $\mathbf{u}$, and $\bar{u} = \sum_{j=1}^{k} u_j / k$.

In constraint (8), the quantity $\sum_{j=1}^{k} (u_j - \bar{u})^2$ measures the variation of summary $\mathbf{y}$'s subtopics degree of coverage. An unbalanced coverage of the subtopics results in a large $\sum_{j=1}^{k} (u_j - \bar{u})^2$, which in turn leads to a relatively low score of $\mathbf{w}^T\Psi(\mathbf{x}, \mathbf{y})$.

4.1.5 Combined Optimization Problem

Considering the three types of constraints, we propose to train a summarization model enforcing diversity, coverage and balance through the following optimization problem:

Optimization Problem 2.

$$\min_{\mathbf{w}, \xi \geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i, \quad (9)$$

subjected to:

$\forall i, \forall \mathbf{y} \in \mathcal{Y} \setminus \mathbf{y}^{(i)} : \xi_i \geq 0,$

1) $\mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}) + \Delta(\mathbf{y}^{(i)}, \mathbf{y}) - \xi_i,$

2) $\mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \sum_{y \in \mathbf{y}^{(i)}} \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, y) - \mu + \xi_i,$

3) $\mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}) + 1 - \| \sum_{y_i \in \mathbf{y}} \mathbf{v}_i \| - \xi_i,$

4) $\mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \geq \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}) + \sum_{j=1}^{k} (u_j - \bar{u})^2 - \xi_i.$

5. SUMMARIZATION THROUGH STRUCTURE LEARNING

The space $\mathcal{Y}$ of all possible subsets is complex. In the following we construct an independence graph to guide our selection of $\mathbf{y} \in \mathcal{Y}$. We then solve the optimization problem (9) following the general cutting plane algorithm [28, 31], and also utilizing the independent graph.

5.1 Independence Graphs

The exploration of the whole space of $\mathcal{Y}$ in either the training or predicting process is known to be intractable. So, the summarization task will greatly benefit from limiting the output space to the most probable subspace. We achieve this by building an *independence graph* for a document.

Given a set of sentences $S = \{s_1, s_2, \ldots, s_k\}$, where s_i represents the i-th sentence in the document, each sentence is considered as a node in the independence graph. There is an edge between s_i and s_j if their similarity is below a predefined threshold η. Specifically, the nodes s_i and s_j with $i < j$ are connected if and only if $\text{sim}(s_i, s_j) < \eta$, where

$$\text{sim}(s_i, s_j) = \frac{m}{\log(|s_i|) + \log(|s_j|)}.$$

Here m is the number of times the same word appears in both sentences. By making use of the paths in the independence graph, we shrink the searching space by avoiding the extraction of two similar sentences as we will demonstrate below.

5.2 Learning Algorithm

In order to solve the optimization problem defined in Equation (9), we employed the cutting plane algorithm [28, 31]. It iteratively adds constraints until the problem has been solved with a desired tolerance ϵ. We start with a group of empty working sets $\mathbf{y}_i$, $\mathbf{y}_i'$, $\mathbf{y}_i''$, for $i = 1, \ldots, n$. Then, we iteratively find the most violated constraints $\bar{\mathbf{y}}$, $\bar{\mathbf{y}}'$, $\bar{\mathbf{y}}''$

Algorithm 1 Cutting plane algorithm

Input $(\mathbf{x}^{(1)}, \mathbf{y}^{(1)}), \ldots, (\mathbf{x}^{(n)}, \mathbf{y}^{(n)}), c > 0, \epsilon > 0$
$\mathbf{y}_i = \emptyset, \mathbf{y}_i^{'} = \emptyset, \mathbf{y}_i^{''} = \emptyset$ for $i = 1, \ldots, n$
repeat
  for $i = 1, 2, \ldots, n$ **do**
    $\omega \equiv \mathbf{w}^T(\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) - \Psi(\mathbf{x}^{(i)}, \mathbf{y}))$
    $\omega^{'} = \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) - \sum_{y \in \mathbf{y}^{(i)}} \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, y) - \mu$
    $H(\mathbf{y}) = \Delta(\mathbf{y}^{(i)}, \mathbf{y}) - \omega$
    $H^{'}(\mathbf{y}) = (1 - \| \sum_{y_i \in \mathbf{y}} \mathbf{v}_i \|) - \omega$
    $H^{''}(\mathbf{y}) = \sum_{j=1}^{k} (u_j - \bar{u})^2 - \omega$
    Compute: $\bar{\mathbf{y}} = \text{argmax}_{\mathbf{y}} H(\mathbf{y})$,
    $\bar{\mathbf{y}}^{'} = \text{argmax}_{\mathbf{y}} H^{'}(\mathbf{y})$, $\bar{\mathbf{y}}^{''} = \text{argmax}_{\mathbf{y}} H^{''}(\mathbf{y})$
    Compute actual slack: $\xi_i = \max\{0, \max_{\mathbf{y} \in \mathcal{W}_i} H(\mathbf{y}), \max_{\mathbf{y} \in \mathbf{y}_i^{'}} H^{'}(\mathbf{y}), \max_{\mathbf{y} \in \mathbf{y}_i^{''}} H^{''}(\mathbf{y})\}$
    if $(H(\bar{\mathbf{y}}) > \xi_i + \epsilon)$ **or** $\left(\omega^{'} < \xi_i + \epsilon\right)$ **or** $\left(H^{'}(\bar{\mathbf{y}}) > \xi_i + \epsilon\right)$ **or** $\left(H^{''}(\bar{\mathbf{y}}) > \xi_i + \epsilon\right)$ **then**
      Add constraint to working set $\mathbf{y}_i \leftarrow \mathbf{y}_i \cup \{\bar{\mathbf{y}}\}$, $\mathbf{y}_i^{'} \leftarrow \mathbf{y}_i^{'} \cup \{\bar{\mathbf{y}}^{'}\}$, $\mathbf{y}_i^{''} \leftarrow \mathbf{y}_i^{''} \cup \{\bar{\mathbf{y}}^{''}\}$
      $\mathbf{w} \leftarrow$ Optimize over $\cup_i (\mathbf{y}_i \cup \mathbf{y}_i^{'} \cup \mathbf{y}_i^{''})$
    end if
  end for
until no working set has changed during iteration.

Algorithm 2 Greedy algorithm using the independence graph

Input:$\mathbf{x}, \mathbf{y}$
Initialize prediction $\bar{\mathbf{y}} \leftarrow \emptyset$
for $i = 1, 2, \ldots, k$ **do**
  $y \leftarrow \text{argmax}_{y \notin \bar{\mathbf{y}}} \mathcal{P}(\mathbf{x}, \mathbf{y}, \bar{\mathbf{y}} \cup \{y\})$, where $\bar{\mathbf{y}} \cup \{y\}$ forms a path in the independence graph
  $\bar{\mathbf{y}} \leftarrow \bar{\mathbf{y}} \cup \{y\}$
end for
return $\bar{\mathbf{y}}$

for each $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ corresponding to the three constraints 1), 3) and 4) in Equation (9), respectively. They are then added to the corresponding working sets, and $\mathbf{w}$ is updated with respect to the new combined working set. The learning algorithm is presented in Algorithm 1. It is guaranteed to halt within a polynomial number of iterations [28].

For each iteration, we need to solve

$$\underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}}\, \mathcal{P}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}, \mathbf{y}) \equiv \underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}}\, \Omega(\mathbf{y}^{(i)}, \mathbf{y}) + \mathbf{w}^T\Psi(\mathbf{x}^{(i)}, \mathbf{y})$$

for $\Omega(\mathbf{y}^{(i)}, \mathbf{y}) = \Delta(\mathbf{y}^{(i)}, \mathbf{y})$, or $(1 - \| \sum_{y_i \in \mathbf{y}} \mathbf{v}_i \|)$, or $\sum_{j=1}^{k} (u_j - \bar{u})^2$. A greedy algorithm using the independence graph, described in Algorithm 2, is proposed to solve this problem where we repeatedly select the sentence y satisfying the following condition: $\bar{\mathbf{y}} \cup \{y\}$ is the sentence set having the highest score while its corresponding node set forms a path in the independence graph. The algorithm ends with an extracted sentence set of size k. This algorithm has the same approximation bound as the greedy algorithm proposed by Khuller et al. [14] to solve the budgeted maximum coverage problem, that is to say, a $(1 - \frac{1}{e})$-approximation bound. So Algorithm 1 has a polynomial time complexity overall.

5.3 Making Prediction

According to (1) and (2), we predict the summary for a given document $\mathbf{x}$ using:

$$\underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}}\, \mathcal{P}(\mathbf{x}, \emptyset, \mathbf{y}) \equiv \underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}}\, \mathcal{F}(\mathbf{x}, \mathbf{y}) = \underset{\mathbf{y} \in \mathcal{Y}}{\text{argmax}}\, \mathbf{w}^T\Psi(\mathbf{x}, \mathbf{y}),$$

which is a special case that can be efficiently solved by Algorithm 2.

5.4 Feature Space

We discuss several features we use for the summarization task. For a document $\mathbf{x}$ and a sentence set $\mathbf{y}$, each component of the feature vector $\Psi(\mathbf{x}, \mathbf{y})$ is set to 1 if the corresponding feature holds true, and 0 otherwise. We first generate the feature vector for each types of features, and then concatenate them to construct the whole feature vector $\Psi(\mathbf{x}, \mathbf{y})$.

5.4.1 Basic Features

These basic features are the most widely used features in summarization, and can be readily computed.

Word Frequency: Word frequency can be used to generate a set of features [31]. To denote the coverage degree of a document for a certain word and the importance of a certain word to a document, two kinds of levels are defined:

Term-importance: It defines the percentage of a certain word in a sentence. For example, the level 0.1 indicates that the frequency of a certain word in some sentence exceeds 10%. Moreover, a separate level is employed to denote the *key words* in the document. A key word is defined by certain rules, which will be discussed later.

Term-extent: It denotes the percentage of the sentences containing a certain word. For example, the level 0.1 denotes that a certain word appears in no less than 10% of the sentences in the document.

For term-importance level A and term-extent level B, there are three binary features:

A word in the term-importance level 0 (appears at least once in a sentence) appears in at least a B fraction of sentences at a term-importance level A.

A word in the term-importance level A appears in at least a B fraction of sentences at a term-importance level 0.

A word in the term-importance level A appears in at least a B fraction of sentences at a term-importance level A.

Let TI be the total number of term-importance levels, and TE be the total number of term-extent levels. A total of $\text{T} = (3 * \text{TI} - 2) * \text{TE}$ features can be then obtained. For each word, we can get a feature vector of length T. The vectors of all words are summed to get a final feature vector.

Position: Every word of the first sentence of a document is labeled as a key word.

Thematic Word: Thematic words are defined as the words with the highest frequency after having deleted the words in the stop-word list. Each thematic word is labeled as a key word.

Length: The number of words in each sentence after removing the words in the stop-word list. We prefer sentences with a length located in a certain range. Several length levels are introduced to denote the length of a sentence. The a feature vector can be generated with each component denotes a length level.

Upper Case words: Each upper case word is labeled as a key word.

5.4.2 Complex Features

We also consider several complex features that prove to be of great help for summarization. Though requiring considerable efforts to generate, the numerous variants of these features are very helpful.

PageRank: The PageRank value of each sentence is calculated in a recursive manner as follows:

$$\mathrm{PR}(s_i) = (1-d)*\mathrm{PR}(s_i)+\sum d*(\mathrm{PR}(s_j)*\mathrm{sim}(s_i,s_j)) \quad (10)$$

where s_i is the i-th sentence in the document, $\mathrm{PR}(s_i)$ is the PageRank value of s_i and d is damper factor, which is set to be 0.85 in our experiments.[2]

There are two kinds of PageRank values: 1) *innPageRank:* Only the similarity between sentences in the same document is calculated. 2) *interPageRank:* Only the similarity between sentences from different documents under the same theme is calculated, other similarity scores are set to zero.

Finally the PageRank value of the sentence s_i is treated as a special key word with a word frequency $\mathrm{PR}(s_i)$.

Two-grams: Two successive words in the same sentence are regarded as a phase. To generate this feature, each phase is denoted as a *bi-word* and defined by an added special importance level of word frequency.

6. EXPERIMENTS

In the experimental study, we aim at addressing the following issues:

1. Does our approach outperform other state-of-the-art approaches? The results demonstrate that it achieves a remarkable improvement over those approaches.

2. Does the incorporation of diversity, coverage, and balance in our framework improves the performance of the summarization? According to our experiments, our proposed constraints do enhance diversity, coverage, and balance of the summary as well as improve performance.

3. Which of the three requirements we deal with in the summarization leads to the best performance? A detailed analysis will be proposed in a later section.

6.1 Data Set

The DUC2001 data set is used for evaluation in our experiments. The data set, denoted as Bigset, contains around 147 summary-document pairs. For this data set, there are mainly two subsets for the summarization task, denoted as Docset1 and Docset2. The respective ground-truth summaries are generated by manually extracting a certain number of sentences from each single document. In the data set there are several themes, and under each theme there are several documents, which enables the *interPageRank* calculation.

For our summarization task, each document set is split into a training data set and a testing data set. A 10-fold cross validation process is employed in the experiments to account for the uncertainty in the data set partition. That is to say, the whole data set is divided evenly into ten folds. Then 9 folds are used for training while the other one is kept for testing.

We use the following preprocessing steps: 1) We eliminate stop words from the original documents and execute stemming using the Porter's stemmer. 2) We calculate each word's frequency and tag each key word defined according to the previous sections. Other features are also calculated and represented in the input files under various forms.

[2]The iteration ends when the scores differ from the last iteration by less than a given threshold (0.001 in our experiment).

6.2 Performance Evaluation

6.2.1 F_1 Evaluation

F_1 measurement is widely used in summarization evaluation. In F_1 evaluation, the predicted summary $\bar{\mathbf{y}}$ and the ground-truth summary $\mathbf{y}$ are compared directly and the precision, recall, F_1 scores are calculated as follows:

$$\Delta(\mathbf{y},\bar{\mathbf{y}}) = \frac{2pr}{p+r}, \quad p = \frac{<\mathbf{y},\bar{\mathbf{y}}>}{<\bar{\mathbf{y}},\bar{\mathbf{y}}>}, \quad r = \frac{<\mathbf{y},\bar{\mathbf{y}}>}{<\mathbf{y},\mathbf{y}>}.$$

6.2.2 ROUGE Evaluation

The ROUGE measure [18] is widely used for evaluation. In fact, the DUC contests usually employ ROUGE measures for automatic summarization evaluation. In ROUGE evaluation, the summarization quality is measured by counting the number of overlapping units, such as n-gram, word sequences, and word pairs between the predicted summary $\bar{\mathbf{y}}$ and the ground-truth summary $\mathbf{y}$. There are several kinds of ROUGE metrics, of which the most important one is ROUGE-N which contains three sub-metrics:

1. ROUGE-N-R is an n-gram recall metric calculated as follows:

$$\text{ROUGE-N-R} = \frac{\sum_{y\in\mathbf{y}}\sum_{\mathrm{gram_n}\in\mathrm{y}}\mathrm{Count_{match}}(\mathrm{gram_n})}{\sum_{y\in\mathbf{y}}\sum_{\mathrm{gram_n}\in\mathrm{y}}\mathrm{Count_{ground}}(\mathrm{gram_n})}$$

2. ROUGE-N-P is an n-gram precision metric calculated as follows:

$$\text{ROUGE-N-P} = \frac{\sum_{y\in\bar{\mathbf{y}}}\sum_{\mathrm{gram_n}\in\mathrm{y}}\mathrm{Count_{match}}(\mathrm{gram_n})}{\sum_{y\in\bar{\mathbf{y}}}\sum_{\mathrm{gram_n}\in\mathrm{y}}\mathrm{Count_{pred}}(\mathrm{gram_n})}$$

3. ROUGE-N-F is an n-gram F_1 metric calculated as follows:

$$\text{ROUGE-N-F} = \frac{2*\text{ROUGE-N-R}*\text{ROUGE-N-P}}{\text{ROUGE-N-R}+\text{ROUGE-N-P}}$$

Here n denotes the length of the n-gram, and $\mathrm{gram}_n \in y$ denotes the n-grams in the document y. $\mathrm{Count_{match}}(\mathrm{gram_n})$ is the number of $\mathrm{gram_n}$ co-occurring in the predicted summary $\bar{\mathbf{y}}$ and the ground-truth summary $\mathbf{y}$, $\mathrm{Count_{ground}}(\mathrm{gram_n})$ represents the occurrence number of $\mathrm{gram_n}$ in the ground-truth summary $\mathbf{y}$, and $\mathrm{Count_{pred}}(\mathrm{gram_n})$ represents the occurrence number of $\mathrm{gram_n}$ in the predicted summary $\bar{\mathbf{y}}$.

According to Lin and Hovy [18], among all the evaluation measures in ROUGE, ROUGE-1 and ROUGE-2 fit the single document summarization evaluation task very well. As a result, for simplicity, in our experiment, only three ROUGE metrics are employed: ROUGE-1, ROUGE-2, ROUGE-W, where ROUGE-W is based on the weighted longest common subsequence. The weight W is set to be 1.2 in our experiments.

6.3 Overall Performance

In this section, several baselines, including both supervised and unsupervised methods are introduced for comparison. This series of experiments is conducted on Docset1, Docset2, and Bigset. To give a concise comparison, only F_1 and ROUGE-2-R are employed for evaluation.

6.3.1 Algorithms for Comparison

Among the supervised methods, we choose Support Vector Machine (SVM) (not structural SVM), Naive Bayes (NB), Logistic Regression (LR), Hidden Markov Model (HMM) and Conditional Random Field (CRF):

SVM: SVM is widely used as a binary classifier, which is appropriate to distinguish summary sentences from non-summary sentences.

NB: This is a approach to classify single class variables in dependence of several feature values. This model follows the assumption that the input variables are conditionally independent.

LR: LR can be regarded as a discriminative version of NB. It is employed to model the posterior probabilities of k classes via linear functions in output variables.

HMM: HMM is an extension to NB for sequentially structured data also representing the dependencies of the input and output as a joint probability distribution.

CRF: CRF differs from HMM by proposing a conditional probability model instead of a joint probability model.

We also compare with four unsupervised methods:

Random: A summary is generated by selecting sentences randomly from a given document.

LEAD: This is a popular baseline on DUC2001 data set. It works by selecting the lead sentences as the summary.

LSA: We identify semantically important sentences using the Latent Semantic Analysis technique and select the k most important sentences as the summary.

HITS: One of the Mihalcea's [20] algorithms based on the authority score of HITS on the directed backward graph. The sentences with the highest authority score are selected in the summary generation.

The results of the baselines' performance on the Bigset are reported in [26].

Our approach is denoted as IndStr-SVM. In our framework, several models can be trained by adapting different strategies in adding constraints and tuning parameters. We set $\mu = 0.5$ and $\eta = 0.4$ in our experiments.

6.3.2 Result and Analysis

Our first experiment is conducted based on only the basic features. Table 1 gives the performance of the unsupervised methods for this experiment, while Table 2 presents the results for the supervised methods.

As shown in Table 1 and Table 2, The random algorithm gives the worst performance as expected. The performance of LEAD is much better than the Random algorithm. LSA achieves better performance than the LEAD algorithm. The result of HITS is the best among the unsupervised approaches, which shows the close relationship between summarization and the graph structure constructed from the sentences.

Compared with those unsupervised methods, supervised methods are generally better with the exception of HITS. As extensions of NB, LR and HMM both result in a remarkable improvement over the performance of NB. SVM proves to be a effective algorithm by achieving a performance similar to that of LR and HMM. CRF takes a big step forward as it achieves much better performance. The comparison between CRF and HMM informs us that CRF does a better job in exploring the dependent structures. Our method obtains the best performance by achieving an increase of 2.0% and 0.3% over CRF on the Bigset in terms of ROUGE-2-R and F_1 respectively and outperforming HITS by 7.4% and 6.0% on the Bigset in term of ROUGE-2-R and F_1 respectively. This can be attributed to the fact that the employment of the three constraints enforcing diversity, coverage, and the balance in the summary.

We also notice that Docset2 is much harder for summarization than Docset1, and our approach makes a greater improvement over other approaches on this task. This demonstrates the robustness of our approach. According to the comparison of the performance on Docset1, Docset2 and Bigset, the gap in the performance between our approach and other approaches is larger on small document sets. This demonstrates that our approach performs better with less training data. This phenomenon may due to the fact that in the generation of a good summarization model, the involvement of the diversity, coverage, and balance issues can partly replace the role played by a large training data set.

Now we discuss experiments using both the basic and complex features. To save space, we presents the results in Tables 3 for the performance of the supervised methods only.

According to Table 3, our approach is still the best while CRF again achieves the second best performance. This again illustrates that the dependence between sentences is important to summarization. Compared with the best unsupervised method HITS, our approach achieves an improvement of 23.7% measured by ROUGE-2-R and 16.3% measured by F_1 on the Bigset. Our approach also outperforms the CRF methods by 10.4% and 2.1% on the Bigset in terms of ROUGE-2-R and F_1, respectively, which is a much larger improvement than when only considering basic features. It illustrates that our approach works better with complex features, i.e., it has larger capacity. Our approach is still robust when using all the features as the gap in the performance between our approach and other approaches is still larger on Docset2.

The robustness of our approach illustrates that the diversity, coverage, and balance issues is important to a summary. We believe that those issues provide the just experience that a good summarization model should learn, and reduce the model's dependence on the training data.

6.4 Strategy Selection

To identity how diversity, coverage, and balance enhance the quality of a summary, we provide several models through a strategy selection based on our approach. Generally speaking, strategies can be sorted into two categories: constraint selection and parameter tuning. Each time, we tune one strategy while others are fixed. This series of experiments is conducted on the Bigset, and the performance is measured in terms of ROUGE-1, ROUGE-2, and ROUGE-W.

6.4.1 Constraint Selection

To understand the effect of each proposed constraint, a series of experiments is conducted by employing different sets

of constraints while training. The comparison is again divided into those based on the basic features and those based on all the features. We denote Indstr-SVM as the model trained with no constraint, Indstr-SVM-C1 as the model trained with the diversity-biased constraint involved, Indstr-SVM-C2 as the model trained with the coverage-biased constraint involved, Indstr-SVM-C3 as the model trained with the balance-biased constraint involved, and Indstr-SVM-All as the model trained with all three constraints involved.

According to Table 4, when using only the basic features, different constraints lead to various degrees of improvement. Indstr-SVM-C2 achieves the best performance among the models trained with a single constraint as it outperforms Indstr-SVM by 4.5% and 4.1% in terms of ROUGE-2-R and ROUGE-2-F respectively. This emphasizes the importance of the coverage issue in summarization. Indstr-SVM-C1 outperforms Indstr-SVM by 4.4% and 3.4% while Indstr-SVM-C3 outperforms Indstr-SVM by 4.2% and 4.0% in terms of ROUGE-2-R and ROUGE-2-F respectively. Indstr-SVM-All achieves the best performance as it outperforms Indstr-SVM by 4.7% and 4.7% respectively.

Table 5 presents us the performance on different constraint set using all the features: Indstr-SVM-C2 again results in the best performance by increasing 3.6% in terms of ROUGE-2-R and 4.0% in term of ROUGE-2-F over Indstr-SVM. For the other two constraints, Indstr-SVM-C1 achieves an improvement of 1.0% and 1.2% respectively, as measured by ROUGE-2-R and ROUGE-2-F compared with Indstr-SVM. Indstr-SVM-C3 outperforms Indstr-SVM by 2.3% and 2.4% in terms of ROUGE-2-R and ROUGE-2-F respectively. Indstr-SVM-All is still the best model, and outperforms Indstr-SVM by 3.9% and 4.4% respectively.

According to the above results, it can be concluded that the coverage-biased constraint makes the greatest contribution to summarization. The result can be explained that the coverage requirement is more important than the other two in the summarization task. When adding all three constraints, a robust model with the best performance of the summarization task is learned. We also notice that there is a little overlap among the effects of those constraints as Indstr-SVM-All just outperforms the models trained with single constraint.

6.4.2 *Parameter Tuning*

The key parameter in our framework is the c used in (9), which aims at keeping a balance between weights and slacks. Figure 1 shows the performance of Indstr-SVM-All based on basic features with the variation of c. The results in Figure 1 also show that the robust model performs better when c is small. This indicates that our constraints focus on weight modification rather than on low training loss.

Figure 2 shows the performance of the robust model based on all features when the value of c varies. According to the results in Figure 2, although there is a few ups and downs as the c parameter varies, the model trained with a small value of c performs relatively better. This shows the same trend as when only considering basic features, which gives another proof that the constraints concerning the diversity, coverage, and balance issues play an important role in summarization.

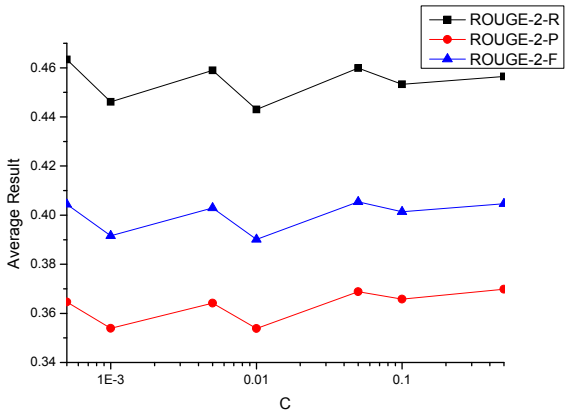

Figure 1: Performance of robust model based on basic features when varying the c value

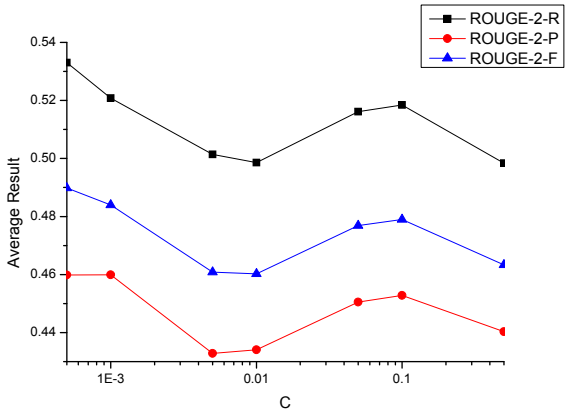

Figure 2: Performance of robust model based on all features when varying the c value

7. CONCLUSION AND FUTURE WORK

In this paper, a novel approach is proposed to train a robust model for document summarization. As a good summary heavily depends on the diversity, coverage, and balance issues, our structure learning approach employs a structural SVM with several constraints to enforce them in a summary. In our approach, we first build an independence graph to capture the structure of the output variable. Then we employ a cutting plane algorithm to solve our proposed optimization problem. Finally, the model obtained in the training process is used to predict the summary when given a new document. Experimental results on the DUC2001 data set demonstrate the good effectiveness of our approach. The performance of our method achieves a remarkable improvement over a set of state-of-art supervised and unsupervised methods and the involvment of diversity, coverage, and balance in summarization proves to be of great help.

In future work, we plan to explore how to enforce diversity, coverage, and balance through feature generation. We will also extend our framework to several applications, including web page summarization and snippets generation for which we will need to modify our method for query-biased summarization. In web page summarization task, there will

Table 1: Results of unsupervised approaches based on basic features

	Docset1	Docset1	Docset2	Docset2	Bigset	Bigset
	ROUGE-2-R	F_1	ROUGE-2-R	F_1	ROUGE-2-R	F_1
RANDOM	0.258	0.252	0.210	0.194	0.245	0.202
LEAD	0.408	0.392	0.216	0.192	0.377	0.311
LSA	0.412	0.340	0.334	0.214	0.382	0.324
HITS	0.453	0.377	0.335	0.239	0.431	0.368

Table 2: Results of supervised approaches based on basic features

	Docset1	Docset1	Docset2	Docset2	Bigset	Bigset
	ROUGE-2-R	F_1	ROUGE-2-R	F_1	ROUGE-2-R	F_1
NB	0.435	0.324	0.312	0.225	0.394	0.336
LR	0.442	0.346	0.346	0.252	0.415	0.349
SVM	0.441	0.348	0.344	0.251	0.416	0.343
HMM	0.441	0.343	0.338	0.243	0.419	0.350
CRF	0.467	0.353	0.360	0.267	0.454	0.389
IndStr-SVM	0.540	0.417	0.467	0.328	0.463	0.390

Table 3: Results of supervised approaches based on all features

	Docset1	Docset1	Docset2	Docset2	Bigset	Bigset
	ROUGE-2-R	F_1	ROUGE-2-R	F_1	ROUGE-2-R	F_1
NB	0.446	0.377	0.383	0.309	0.436	0.372
LR	0.463	0.395	0.401	0.329	0.450	0.383
SVM	0.468	0.397	0.407	0.335	0.449	0.385
HMM	0.461	0.380	0.395	0.328	0.451	0.380
CRF	0.495	0.420	0.416	0.340	0.483	0.419
IndStr-SVM	0.574	0.475	0.519	0.390	0.533	0.428

Table 4: Performance on different constraint set based on basic features

constraint set	Indstr-SVM	Indstr-SVM-C1	Indstr-SVM-C2	Indstr-SVM-C3	Indstr-SVM-All
ROUGE-1-R	0.58354	0.59257	0.59939	0.59682	0.59569
ROUGE-1-P	0.45801	0.46374	0.46959	0.46906	0.46928
ROUGE-1-F	0.50773	0.51344	0.52034	0.51881	0.51902
ROUGE-2-R	0.43925	0.45843	0.45912	0.45758	0.45990
ROUGE-2-P	0.35219	0.36466	0.36658	0.36637	0.36880
ROUGE-2-F	0.38733	0.40040	0.40337	0.40267	0.40543
ROUGE-W-R	0.23701	0.24293	0.24762	0.24641	0.24466
ROUGE-W-P	0.32515	0.33348	0.33905	0.33841	0.33693
ROUGE-W-F	0.27045	0.27862	0.28210	0.28088	0.27951

Table 5: Performance on different constraint set based on all features

constraint set	Indstr-SVM	Indstr-SVM-C1	Indstr-SVM-C2	Indstr-SVM-C3	Indstr-SVM-All
ROUGE-1-R	0.60669	0.60854	0.62368	0.61264	0.62121
ROUGE-1-P	0.52589	0.52988	0.54068	0.53254	0.54254
ROUGE-1-F	0.55725	0.56094	0.57408	0.56437	0.57354
ROUGE-2-R	0.50104	0.50587	0.51892	0.51274	0.52081
ROUGE-2-P	0.43894	0.44420	0.45654	0.44874	0.45994
ROUGE-2-F	0.46351	0.46897	0.48194	0.47464	0.48401
ROUGE-W-R	0.26051	0.26235	0.26899	0.26353	0.26995
ROUGE-W-P	0.39510	0.39920	0.40855	0.39972	0.41173
ROUGE-W-F	0.30994	0.31300	0.32080	0.31398	0.32214

be other types of information available, such as hyperlink for summary structure generation, which is expected to lead to better performance.

8. ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable and constructive comments. Gui-Rong Xue thanks National Natural Science Foundation of China (NO. 60873211) for such generous support. Part of the work of the fourth author is supported by the 111 Project (Grant No. B07022) jointly funded by Ministry of Education and State Administration of Foreign Experts Affairs, PRC.

9. REFERENCES

[1] Ibm many aspects document summarization tool, http://www.alphaworks.ibm.com/tech/manyaspects. .

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *WWW*, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *SIGIR*, pages 335–336, New York, NY, USA, 1998. ACM.

[4] C. L. A. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pages 659–666, New York, NY, USA, 2008. ACM.

[5] J. M. Conroy and D. P. Oąrleary. Text summarization via hidden markov models. In *SIGIR*, pages 406–407, New York, NY, USA, 2001. ACM.

[6] M. desJardins, E. Eaton, and K. Wagstaff. Learning user preferences for sets of objects. In *ICML*, pages 273–280, New York, NY, USA, 2006. ACM.

[7] G. ErKan and D. R. Radev. Lexpagerank: Prestige in multi-document text summarization. In *EMNLP*, Barcelona, Spain, 2004.

[8] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* newblock 2001.

[9] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *NAACL-ANLP*, pages 40–48, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[10] Y. H. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *SIGIR*, pages 19–25, New York, NY, USA, 2001. ACM.

[11] S. Harabagiu and F. Lacatusu. Topic themes for multi-document summarization. In *SIGIR*, pages 202–209, New York, NY, USA, 2005. ACM.

[12] H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. B. Wise, and X. Zhang. Cross-document summarization by concept classification. In *SIGIR*, pages 121–128, New York, NY, USA, 2002. ACM.

[13] H. Jing and K. R. McKeown. Cut and paste based text summarization. In *ANLP*, pages 178–185, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[14] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[16] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.

[17] J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *SIGIR*, pages 68–73, New York, NY, USA, 1995. ACM.

[18] C. Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *NAACL*, pages 71–78, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[19] D. Metzler and T. Kanungo. Machine learned sentence selection strategies for query-biased summarization. In *SIGIR*, 2008.

[20] R. Mihalcea. Language independent extractive summarization. In *AAAI*, pages 1688–1689, 2005.

[21] R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In *EMNLP*, Barcelona, Spain, 2004.

[22] M. Naaman and L. Kennedy. Generating diverse and representative image search results for landmarks. In *WWW*, pages 297–306, New York, NY, USA, 2008. ACM.

[23] T. Nomoto and Y. Matsumoto. A new approach to unsupervised text summarization. In *SIGIR*, pages 26–34, New York, NY, USA, 2001. ACM.

[24] C. V. Rijsbergen. *Information Retrieval.* 1979.

[25] D. Shen, Z. Chen, Q. Yang, H. J. Zeng, B. Zhang, Y. Lu, and W. Y. Ma. Web-page classification through summarization. In *SIGIR*, pages 242–249, New York, NY, USA, 2004. ACM.

[26] D. Shen, J. T. Sun, H. Li, Q. Yang, and Z. Chen. Document summarization using conditional random fields. In *IJCAI*, pages 2862–2867, 2007.

[27] J. T. Sun, D. Shen, H. J. Zeng, Q. Yang, Y. C. Lu, and Z. Chen. Web-page summarization using clickthrough data. In *SIGIR*, pages 194–201, New York, NY, USA, 2005. ACM.

[28] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

[29] K. Wagstaff, M. desJardins, E. Eaton, and J. Montminy. Learning and visualizing user preferences over sets. In *AAAI*, 2007.

[30] X. Wan, J. Yang, and J. Xiao. Collabsum: Exploiting multiple document clustering for collaborative single document summarizations. In *SIGIR*, pages 143–150, New York, NY, USA, 2007. ACM.

[31] Y. Yue and T. Joachims. Predicting diverse subsets using structural svms. In *ICML*, pages 1224–1231, New York, NY, USA, 2008. ACM.

[32] H. Y. Zha. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *SIGIR*, pages 113–120, New York, NY, USA, 2002. ACM.