

Comparative Document Summarization via Discriminative Sentence Selection

Dingding Wang[†] Shenghuo Zhu[‡] Tao Li[†] Yihong Gong[‡]
[†] School of Computer Science
Florida International University
Miami, FL 33199
[†] {dwang003,taoli}@cs.fiu.edu
[‡] NEC Laboratories, America, Inc.
10080 N. Wolfe Rd. SW3-350
Cupertino, CA 95014
[‡]{zsh,ygong}@sv.nec-labs.com

ABSTRACT

Given a collection of document groups, a quick question is what are the differences in these groups. In this paper, we study a novel problem of summarizing the differences between document groups. A discriminative sentence selection method is proposed to extract the most discriminative sentences which represent the specific characteristics of each document group. Experiments on real world data sets demonstrate the effectiveness of our proposed method.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; I.2.6 [Artificial Intelligence]: Learning

General Terms

Algorithms, Experimentation, Performance

Keywords

Comparative Document Summarization, Discriminative Sentence Selection

1. INTRODUCTION

In many applications, when facing a set of document groups sharing similar topics, people are interested to know the differences in these document groups. Thus a summary describing major differences among the given documents is necessary to facilitate the comparison of these document groups, e.g., summarizing different points of view in news articles, comparing different blog communities, and summarizing the changes in the community evolution. To the best of our knowledge, the problem of summarizing the distinctness of documents has not been well defined and studied. Thus in this paper, we study the novel problem referred to as *Comparative Extractive Document Summarization* (CDS) to summarize the differences between comparable document

groups. Specifically, given a collection of document groups, the CDS problem is to generate a short summary delivering the differences of these documents by extracting the most discriminative sentences in each document group. This problem is related to the traditional document summarization problem since both of them try to extract sentences from documents to form a summary. However, traditional document summarization aims to cover the majority of information among document collections, while our goal is to find differences.

In this paper, we propose a discriminative sentence selection approach based on a multivariate normal generative model to extract sentences best describing the unique characteristics of each document group. Given a collection of document groups (clusters), we decompose these documents into sentences, and calculate sentence-document and sentence-sentence similarities using cosine similarity. Since each document is labeled to indicate which cluster it belongs to, we select sentences one by one to minimize the average variance of all the cluster targets under the distribution estimation based on a multivariate normal generative model. Experiments on real world data sets demonstrate the effectiveness and the discriminative ability of the summaries generated by our method.

2. RELATED WORK

Traditional document summarization aims to generate a summary delivering the major information expressed in a collection of documents. Current methods usually ranks the sentences in the documents according to the scores calculated by a set of predefined features, such as term frequency-inverse sentence frequency (TF-ISF) [5], sentence or term position [8], and number of keywords [8]. Other techniques such as latent semantic analysis (LSA) [3], matrix factorization [6], hidden Markov model [1], and graph ranking [2] are also used in document summarization.

There are few works focusing on comparing documents. Some work on comparing product reviews has been proposed in [4]. Although the work summarizes and compares the positive/negative aspects of products, the essence of the work is still based on word-level opinion mining. Another work referred to as comparative text mining (CTM) [9] tries to discover common and specific themes in multiple documents using a generative probabilistic mixture model. However, the word-level representation used in the work has limited interpretation ability. There is also very recent work on evolutionary summarization which compares the different event evolutionary phases [7].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM '09, November 2–6, 2009, Hong Kong, China.

Copyright 2009 ACM 978-1-60558-512-3/09/11 ...\$10.00.

3. PROBLEM FORMULATION

Suppose we have f sentences of the document collection, denoted by $\{X_i|i \in F\}$, where F is the full sentence index set, having $|F| = f$. We have the group variable, Y , represented by multiple group indicator variables. The problem of sentence selection is selecting a subset of sentences, $S \subset F$, to accurately discriminate the documents in different groups, i.e. to predict the group identity variable Y , given that the cardinality of S is m ($m < f$). Let us denote $\{X_i|i \in S\}$ by X_S , for any set S . The prediction capability of Y given X_S can be measured by the entropy of Y given X_S , which is defined as

$$H(Y|X_S) \stackrel{\text{def}}{=} -E_{p(Y,X_S)}(\ln p(Y|X_S)), \quad (1)$$

where $E_p(\cdot)$ is the expectation given the distribution p , and p stands for the underlying document distribution, i.e. the joint distribution $p(Y, X_s)$. The sentence selection problem using the mutual information criterion is

$$\arg \min_S H(Y|X_S). \quad (2)$$

4. DISCRIMINATIVE SENTENCE SELECTION

Selecting an optimal subset of sentences is a combinatorial optimization problem, which is an NP-hard problem. The effective practice is to take a greedy approach, i.e., sequentially selecting features to achieve a sub-optimal solution.

4.1 Multivariate Normal Model

We assume that the joint distribution of $\{X_i\}$ and Y is a multivariate normal distribution,

$$\mathbf{z} = \begin{pmatrix} X_F \\ Y \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (3)$$

where $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix. Let F be the index set of X in \mathbf{z} , and T be the index set of Y in \mathbf{z} .

We denote the sentence-document similarity matrix by $\tilde{\mathbf{X}}$, where each row represents a sentence, and each column represents a document. For example, the sentence-document similarity matrix can be constructed using the dot product of the sentence-term and term-document matrices which are computed using cosine similarity. We consider multiple target variables. For grouped documents, we denote the group identity matrix as $\tilde{\mathbf{Y}}$, where each column represents group identity variables of a document, and each row represents a group identity variable.

Given the data, we estimate the parameters of Eq. (3) by

$$\hat{\boldsymbol{\mu}} \stackrel{\text{def}}{=} \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{X}} \\ \tilde{\mathbf{Y}} \end{pmatrix} \mathbf{1}, \quad \hat{\boldsymbol{\Sigma}} \stackrel{\text{def}}{=} \frac{1}{n} \mathbf{Z}\mathbf{Z}^\top, \quad (4)$$

where n is the number of rows of matrix $\tilde{\mathbf{X}}$, $\mathbf{1}$ is a column vector of size n , whose elements are all ones, and

$$\mathbf{z} \stackrel{\text{def}}{=} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \stackrel{\text{def}}{=} \begin{pmatrix} \tilde{\mathbf{X}} \\ \tilde{\mathbf{Y}} \end{pmatrix} - \hat{\boldsymbol{\mu}}\mathbf{1}^\top. \quad (5)$$

Next, we consider the sentence-sentence similarity matrix \mathbf{W} . For example, the sentence-sentence similarity matrix can be obtained by the product of the standardized sentence-term matrix and its transpose. We use sentence-sentence similarity matrix to augment the covariance matrix. Also we

consider to add a regularization term to prevent the ill-posed problem of the estimation. Then, we define our covariance matrix by

$$\boldsymbol{\Sigma} = \hat{\boldsymbol{\Sigma}} + \alpha \begin{pmatrix} \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_t - \frac{1}{t}\mathbf{1}\mathbf{1}^\top \end{pmatrix} + \lambda \begin{pmatrix} \mathbf{I}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_t \end{pmatrix}, \quad (6)$$

where \mathbf{I} is the identity matrix of size of the number of groups, α is a mixture parameter to weigh the importance of the sentence-sentence matrix, λ is the regularization parameter to increase the robustness. The reason for the lower-right corner of the second term is that we consider that the document group are exclusive.

4.2 Sequential Selection Method

In the multivariate normal model, the sentence selection problem in Eq. (2) becomes

$$\arg \min_S \ln |\boldsymbol{\Sigma}_{T|S}|, \quad (7)$$

where $\boldsymbol{\Sigma}_{T|S} \stackrel{\text{def}}{=} \boldsymbol{\Sigma}_{TT} - \boldsymbol{\Sigma}_{TS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{ST}$, known as Schur complement. As the determinant of the covariance matrix is known as *generalized variance*. This criterion is to minimize the generalized variance of the joint distribution of targets.

We use the greedy approach to solve Eq. (7). Let $\mathbf{K} = \boldsymbol{\Sigma}_{D|S}$, where D is the full index set. Based on the property of multivariate normal distribution, we have

$$\begin{aligned} \ln |\boldsymbol{\Sigma}_{T|S \cup \{i\}}| &= \ln \left| \mathbf{K}_{TT} - \frac{1}{K_{ii}} (\mathbf{K}_{Ti}\mathbf{K}_{iT}) \right| \\ &= \ln |\mathbf{K}_{TT}| + \ln \left(1 - \frac{\mathbf{K}_{iT}(\mathbf{K}_{TT})^{-1}\mathbf{K}_{Ti}}{K_{ii}} \right). \end{aligned}$$

Therefore

$$\arg \min_i \ln |\boldsymbol{\Sigma}_{T|S \cup \{i\}}| = \arg \max_i \frac{\mathbf{K}_{iT}(\mathbf{K}_{TT})^{-1}\mathbf{K}_{Ti}}{K_{ii}}.$$

We can compute $\boldsymbol{\Sigma}_{D|S \cup \{i\}}$ from $\mathbf{K} = \boldsymbol{\Sigma}_{D|S}$ by

$$\boldsymbol{\Sigma}_{D|S \cup \{i\}} = \mathbf{K} - \frac{1}{K_{ii}} (\mathbf{K}_{Di}\mathbf{K}_{iD}).$$

Algorithm 1 shows the computational procedure. The procedure is similar to the sequential algorithm in [10].

Algorithm 1 Discriminative Sentence Selection (DSS)

Input: m : number of selected sentences;
 $\boldsymbol{\Sigma}$: obtained from Eq. (6);

Output: S : selected sentences;

- 1: $\mathbf{K} = \boldsymbol{\Sigma}$;
 - 2: $S = \emptyset$;
 - 3: **repeat**
 - 4: $i = \arg \max_{i \notin S} (\mathbf{K}_{iT}(\mathbf{K}_{TT})^{-1}\mathbf{K}_{Ti})/K_{ii}$;
 - 5: $\mathbf{K} \leftarrow \mathbf{K} - (\mathbf{K}_{Di}\mathbf{K}_{iD})/K_{ii}$;
 - 6: $S \leftarrow S \cup \{i\}$;
 - 7: **until** $|S| = m$.
-

5. A CASE STUDY

A case study is conducted to examine the summarization results by different methods on real Blog data. Each method forms a one-sentence summary for each blog community.

5.1 Blog Data

The real blog data was collected by an in-house blog crawler during 2005 and 2006. In this data set, we have 407 English blogs with 274,679 entries in 441 days (63 weeks) between July 10th in 2005 and September 23rd in 2006. The data set contains 7 communities as follows: **1. War and Terrorist:** discusses the war and conflicts with Iraq; **2. Race Issues:** consists of entries on the topic of race and ethnic policy and facts; **3. Duke Lacrosse Case:** describes the scandal that started in March 2006 when a black stripper falsely accused three white members of Duke University’s men’s lacrosse team of raping her; **4. Religion:** mainly focuses on stories about the Christian religion; **5. 911 Commission:** discusses the national commission on terrorist attacks upon the United States on September 11, 2001; **6. China Issues:** contains entries about China’s democracy, politics, and economics; **7. Hurricane Katrina:** describes the destroy of the hurricane Katrina in New Orleans in 2005, and discusses the government’s behavior in this disaster.

5.2 Implemented Systems

We implement the following systems in the case study. **NMF-1:** performs non-negative matrix factorization on the sentence-term matrix. **K-Means:** conducts K-Means clustering on sentences and includes the centroid sentences into the summary. **LSA:** conducts latent semantic analysis on terms by sentences matrix as proposed in [3]. **LeadBase:** returns the leading sentences of all the documents for each topic. **Center:** selects the sentence most similar to all of the rest sentences in each community. **NMF-2:** performs non-negative matrix factorization on the sentence-document matrix. **DSS:** our proposed discriminative sentence selection method.

5.3 Results and Discussion

Discriminative Sentence Selection	
1	There is no cold war, there is no Saddam. Lebanon has also changed.
2	If hiring rap sheet-free intelligent people means they won’t hire a black applicant for another five years.
3	He should drop the case against the lacrosse players but not the sexual assault case itself.
4	Rahman, who is about 41 years old, converted from islam to christianity over 16 years ago.
5	To be totally honest with you, we believed that there may have been a classified annex.
6	I suspect that his position reflects conventional wisdom among the Chinese military establishment.
7	In both the short and long term what those displaced by hurricane Katrina need most is money.

Table 1: Sentences selected by our proposed DSS approach The first column represents the community ID to which the selected sentences belong.

Table 1 and 2 show the sentences selected by different implemented systems. From the results, we observe that the widely used existing document summarization methods such as NMF-1, K-Means and LSA can not extract sentences covering all the seven communities contained in these blog entries. For example, NMF-1 method extracts two sentences from 911 commission and hurricane Katrina communities respectively, however, the duke lacrosse case and war and terrorist communities are missing. Another problem of these methods is that many of the selected sentences can

not discriminate the communities because 1) they come from the same community (for example, there are three sentences from hurricane Katrina in the results of K-Means method, and three sentences from war and terrorist in the results of LSA method); 2) some of the selected sentences are too general, such as “The system is designed to keep you running in circles so you won’t see the real issues” in the results of NMF-1 method. This sentence is supposed to describe the specific content in the race issue community, however, the sentence is not distinctive at all and may appear in any community. The LeadBase and Community Center approaches explore each community one by one, so the sentences they select definitely come from different communities. And it is nice that NMF-2 can automatically extract one sentence from each of the seven communities. However, there are still some sentences selected by these methods are non-discriminative. In Table 2, we label the non-discriminative sentences using a crossing below the community ID.

While looking at the results by our proposed discriminative sentence selection method, each of the sentences represents one community respectively, and the specific characteristics of the community are well summarized. In Table 1, we highlight some keywords representing the unique features of each topic.

6. CONCLUSION

In this paper, a discriminative sentence selection method is proposed to summarize the differences of document groups based on the multivariate normal model. Experiments on real world data show the effectiveness of the proposed method.

Appendix: The work is partially supported by NSF grants IIS-0546280 and DMS-0844513.

7. REFERENCES

- [1] J. Conroy and D. O’Leary. Text summarization via hidden markov models. In *Proceedings of SIGIR*, 2001.
- [2] G. Erkan and D. Radev. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*, 2004.
- [3] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of SIGIR*, 2001.
- [4] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of SIGKDD*, 2004.
- [5] D. Radev, H. Jing, M. Stys, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, pages 919–938, 2004.
- [6] D. Wang, T. Li, S. Zhu, and C. Ding. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of SIGIR*, 2008.
- [7] D. Wang, L. Zheng, T. Li, and Y. Deng. Evolutionary document summarization for disaster management. In *Proceedings of SIGIR*, 2009.
- [8] W.-T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. Multi-document summarization by maximizing informative content-words. In *Proceedings of IJCAI*, 2007.
- [9] C. Zhai, A. Velivelli, and B. Yu. A cross-collection mixture model for comparative text mining. In *Proceedings of SIGKDD*, 2004.
- [10] S. Zhu, D. Wang, K. Yu, T. Li, and Y. Gong. Feature selection for gene expression using model-based entropy. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008.

NMF-1		K-Means		LSA	
7	After watching the media bungle the coverage in New Orleans, I can only imagine what is really happening in Iraq.	4X	He didn't say it was gone; plenty of journalists still heard the call.	1	The column correctly identifies Tehran's diplomatic game: Iran's hezbollah proxies engaging Israel deflects political attention from Iran's nuclear shenanigans.
5	The federal prosecutor's office has had two Iraqi men arrested on Sunday and Tuesday of this week.	3	No one, of course, wants to wait for the evidence except the prosecutor, who hasn't charged anyone.	1	The Lebanese have also experienced twenty years of syrian occupation and thuggery.
2X	The system is designed to keep you running in circles so you won't see the real issues.	6	But in march, Kleinsmith was ordered to cease all work on the project.	1X	It suggested the case but not at the center of its public diplomacy.
6	A snippet: even in a country that celebrates free speech, you don't spontaneously vocalize grievances at state events.	7	S. O'Brien: you're telling me the president told you the governor said she needed 24 hours to make a decision?	2	The first thing whites do when making these decisions is appeal to the lowest element.
5	I discussed this in earlier posts, but it bears repeating: terrorists can change tactics situationally.	7	While the feds were killing themselves to get poor black people out the city, the white people got ignored.	6	If analysts could establish a legitimate reason to investigate a person further, they could keep the corresponding data.
7X	Indeed... update 2 (by kevin): two issues with the story have been raised via comment and trackback.	7X	80% of what the national media has reported thru all this was flawed in some non-trivial way.	4	Purchasing power parity, PPP, is among other things, a way to let poor countries feel better about themselves.
4	Reform efforts have been slow, say experts, since there are so few judges and lawyers with experience.	5	German authorities, acting on CIA recommendations, had been focused on monitoring the activities of islamic groups linked to Bin Ladin.	7	If you're comfortable with that arrangement I'd like to personally encourage you to donate.
LeadBase		Center		NMF-2	
1	The washington post has the best reporting from iraq that I've found.	1	For this, he says, the bush team "needs the dying, withering, but still powerful press axis."	1	A shift to "no proxy war" would put it in a vise between Israel and Iraq.
2X	When I first saw the movie Carmen Jones, I was convinced that back in the day, Harry Belafonte was the finest thing walking.	2	Start with the predominantly black, heavily-funded government school system in our nation's capital.	2	My indignation against this garbage is only one of many reasons black liberals "take issue" with me.
3	The "duke rape" case has been salaciously splashed all over the news.	3	There is no current trend of white men raping black women, in other words.	3	Blacks tend to attack blacks, and whites tend to attack whites.
4	"Things I Used to Teach that I No Longer Believe" was the Title of the Panel ... at the journalism professors' annual convention.	4X	He didn't say it was gone; plenty of journalists still heard the call.	4X	He didn't say it was gone; plenty of journalists still heard the call.
5	One More Look At Prague My Last Post reviews a rather obscure report on the discovery of an Iraqi spy ring in Germany in February or March of 2001, resulting in the capture of two Iraqi Intelligence Services agents.	5	One person comes to the commission a week prior to the release of the report and says, wait.	5	One cannot help but draw the conclusion, especially in this case, that the commission deliberately excluded it from their report.
6	Notes on Chinese first-strike threat Since I think it's probably the most important real story of the day.	6	Chinese sources seem to raise doubts on the official government version of the story.	6	Not going to discuss my or CNN's conversations with the Chinese government, he said.
7	Several people have asked what, if anything, they could do to help.	7	While the feds were killing themselves to get poor black people out the city, the white people got ignored.	7	While the feds were killing themselves to get poor black people out the city, the white people got ignored.

Table 2: Sentences selected by NMF-1, K-Means, LSA, LeadBase, Center, and NMF-2 approaches. The "X" next to the community ID represents that the sentence selected for that community is too general and not discriminative.