

# What Makes a Good Summary?

Qunhua Zhao<sup>1</sup>, Eugene Santos<sup>1</sup>, Hien Nguyen<sup>2</sup>, Ahmed Mohamed<sup>3</sup>

<sup>1</sup> Thayer School of Engineering,  
Dartmouth College, Hanover, NH 03755  
{Qunhua.Zhao,Eugene Santos Jr.}@Dartmouth.edu

<sup>2</sup> Department of Mathematical and Computer Sciences  
University of Wisconsin-Whitewater, WI 53190

<sup>3</sup> Department of Computer Science and Engineering  
University of Connecticut, Storrs, CT 06269

**Abstract.** One of the biggest challenges for intelligence analysts who participate in prevention or response to a terrorism act is to quickly find relevant information from massive amount of data. Along with research on information retrieval and filtering, text summarization is an effective technique to help intelligent analysts shorten their time to find critical information and make timely decision. In this paper, we report our experiment to study the sensitivity of human and multi-document summarization. We use the DUC 2002 collection for multi-document summarization. Two groups of document sets are considered which are the sets consisting of closely correlated documents with highly overlapped content; and the sets of diverse documents covering a wide scope of topics. Intuitively, this suggests that creating a quality summary would be more difficult for the latter case. However, human evaluators were discovered to be fairly insensitive to this difference. This occurred when they were asked to rank the performance of automated summarizers. In this paper, we examine and analyze our experiments in order to better understand this phenomenon and how we might address it to improve summarization quality. In particular, we present a new metric based on document graphs that can distinguish between the two types of document sets.

**Keywords:** Multi-document summarization, user evaluation, ranking correlation

## 1 Introduction

To prevent or quickly respond to a terrorism act, every intelligent analyst needs to gather critical information and makes decision based on retrieved information under time pressure. He/she deals with a huge volume of online and offline information resources available on a daily basis. The requirement for quickly obtaining critical information in order to make timely decisions has further exacerbated this problem. Along with research on search engines, automatic text summarization has been proposed as one natural approach to dealing with this problem [1][5]. Usually, summaries are incorporated into the information retrieval process to provide readers with the key information in the original texts, and to help them judge relevancy of the texts to the tasks at hand so that they can decide quickly whether it is worth going through the full texts.

Text summarizations has been defined as “the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks)” [5]. This definition includes three elements: (1) a summary is a condensed version that contains important information; (2) a summary can be generated for a single document or a set of documents (the latter is called *multi-document summarization*); and, (3) a summary should be user- and task-oriented. Taking into account the possible uses that a summary may serve, it can be indicative by pointing out the topics which have been addressed in the text, or informative by covering as much as possible, important content and/or critical information that offers a critique of the source. Hence, the intention and coverage of summaries can be different depending on tasks. Even for the same tasks, various people may have different opinions on what pieces of information are interesting and important. However, the prevalent form of summaries that we usually encounter is the generic summary that targets a wide range of readers. Until recently, user-sensitive (or user-centered) summarization has become an increasingly active research area which focuses on how individual differences affect a user’s judgment on what should be included in a summary [1].

Intuitively, multi-document summarization seems to be a more difficult task than creating a summary from a single text, given that a set of documents typically cover a variety of topics. For example, assuming that there is a document set containing news reports on the Mumbai commuter rail explosion in India on July 11 2006, the topics included may vary from what has happened in nearby locations, the injuries or deaths, and the terrorism organizations that may involve. As has been pointed out, it is extremely difficult to determine what should be covered in a good summary because of the breadth of the document set [7]. At the same time, multi-document summarization has great potential in assist intelligence analysts in their daily work, where they are likely to receive related messages/reports/documents in groups.

In this paper, we work with the standard testbed from the Document Understanding Conference (DUC) 2002 data collection [12] to find an answer to the question: What is needed for a good summary? Within the DUC 2002 data collection for multi-document summarization, there are basically two groups of document sets: (1) document sets which consist of closely related documents; and (2) those of highly diverse texts. Intuitively, it should be much more difficult to create a good summary for the document sets in the latter case. We conduct a user study to examine if the difference between these two groups of document sets has any impacts on judgments about the qualities of the summaries included in the corpus. It was expected that human evaluators could identify this difference easily. Surprisingly, this is not the case. The impact of this difference on human judgments of summarization quality can only be identified by further detailed analysis. As such, we try to determine if differences between two groups are quantifiable. We find that our document graph (*DG*) approach to measure content is capable of doing so. Thus, it allows us to alert a user to not take the summary at face value.

For the insensitivity exhibited by human evaluators, we believe that this arises from the fact that there were no specific guidelines or tasks required during the summary evaluations. In this situation, we believe that human readers tend to accept summaries which simply contain general information as quality summaries. In this paper, we present and analyze our experiments in order to better understand this

phenomenon, how we address it to improve summarization, and better satisfy users' needs.

In the following text, we first introduce the data collection we used in the experiment. Next, we describe the document graph approach that we used to identify the two groups of document sets. Our experiments on evaluating the performance of various automatic summarization systems (summarizers) are followed and a comparison between different summarization ranking approaches currently in use is provided. Finally, we conclude with a discussion of our results.

## **2. DUC 2002 Data Collection for multi-document summarization.**

The National Institute of Standards and Technology launched a study on automatic text summarization and evaluation called the Document Understanding Conference (DUC). Since 2001, different types of summarization tasks have been studied, such as single document summarization, multi-document summarization, extract and abstract generation, and headline generation. A data collection is provided to serve as a test bed for state-of-the-art algorithms and systems.

DUC 2002 data collection for multi-document summarization was used in our experiments. It has 59 document sets. Each set has from 5 to 15 documents (with an average of 10). This collection contains articles from the Wall Street Journal, AP newswire, San Jose Mercury News, Financial Times, LA Times, and FBIS records **Error! Reference source not found.** The document sets have been classified into four categories:

- Category 1: Documents about a single natural disaster and created within at most a seven day window (*one event, disaster domain, limited time* category).
- Category 2: Documents about a single event in any domain created within at most a seven day window (*one event, any domain, limited time* category).
- Category 3: Documents about multiple distinct events of a single type (no limit on the time window) (*multiple events, unlimited time* category).
- Category 4: Documents that present biographical information mainly about a single individual (*individual biography* category).

There were ten participants that submitted summarization outputs generated by their automatic summarizers. For each document set, two model summaries were also created by human assessors. These summaries are extracts. Two different sizes of extracts, 200 words or 400 words, have been generated for each document set. For this paper, we worked with 200 word extracts in our experiments in order to reduce the time for our user study.

## **3. Using document graph approach to identify different needs for summarization**

Generating a summary is actually a process of extracting important relevant information and then presenting it to the user. Accordingly, we use an approach called document graph (*DG*) generation for information extraction and representation, which is described below, and use the resulting *DG* for the purpose of document similarity comparison (also see [15,16]).

### 3.1 Document Graph (DG)

A document graph (*DG*) is a directed graph of concepts/entities and the relations between them [15,16] It contains two kinds of nodes, concept/entity nodes and relation nodes. Currently, only two kinds of relations, “*isa*” and “*related to*”, are captured for simplicity. The construction of a *DG* is an automated process, which contains following steps, (1) tokenizing a document in the plain text format into sentences (a summary is treated the same as a document); (2) parsing each sentence by using Link Parser [18], (3) extracting noun phrases (NPs) from the parsing results; and, (4) generating relations between concepts/entities based on heuristic rules, and put them into the graph format. The most computationally costly step is parsing the sentence, with a complexity of  $O(m^3)$  where  $m$  is the number of words in a sentence [18]. For graph generation and comparison, we note that we are working strictly with labeled graphs as opposed to general graph isomorphism. We also greatly improve efficiency by using hashing methods.

We use three heuristic rules for relation generation which are:

- The NP-heuristic: it helps set up the hierarchical relations. For example, from a NP “*folk hero stature*”, we generate relations “*folk hero stature - isa - stature*”, “*folk hero stature - related to - folk hero*”, and “*folk hero - isa - hero*”.
- The NP-PP-heuristic: it attaches all prepositional phrases to adjacent NPs. For example, from “*workers at a coal mine*”, we generate a relation, “*worker - related to - coal mine*”.
- The sentence-heuristic: it relates all the concepts/entities contained within one sentence. The relations created by sentence-heuristic are then sensitive to verbs, since the interval between two noun phrases usually contains a verb. For example, from a sentence “*workers at a coal mine went on strike*”, we generate a relation “*worker - related to - strike*”. Another example, from “*The usual cause of heart attacks is a blockage of the coronary arteries*”, we generate “*heart attack cause - related to - coronary artery blockage*”. Figure 1 shows an example of a partial *DG*.

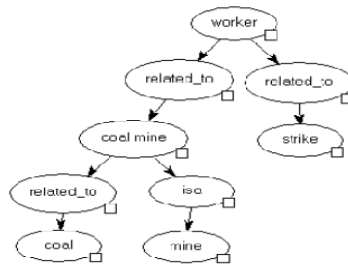


Figure 1. A partial *DG*.

We evaluate the similarity between the two documents based on the *DGs* that are generated from them. The simple similarity of two *DGs*,  $DG_1$  to  $DG_2$ , is given by the equation:

$$Similarity(DG_1, DG_2) = \frac{n}{2N} + \frac{m}{2M} \quad (1)$$

which is modified from Montes-y-Gómez et al, [11]. In the equation,  $N$  is the number of concept/entity nodes in  $DG_1$ , and  $M$  stands for number of relations in  $DG_1$ ;  $n$  is the number of matched concept/entity nodes in two  $DG$ s, and  $m$  is the number of matched relations. Two relation nodes are matched only when at least one of its parent nodes and one of its child nodes are also matched. Since we might compare two  $DG$ s that are significantly different in size (for example,  $DG$ s representing an extract and its source document), we used the number of concept/entity nodes and relation nodes in the target  $DG$  as  $N$  and  $M$ , instead of the total number of nodes in both  $DG$ s.  $Similarity(DG_1, DG_2)$  provides a percentage of  $DG_1$  that is contained in  $DG_2$ ; at the same time,  $Similarity(DG_1, DG_2)$  represents how much of  $DG_2$  has been covered in  $DG_1$ . Next, an F-score can be calculated by Equation (2) [19].

$$F = \frac{2 \times P \times R}{(P + R)} \quad (2)$$

In the equation,  $P$  is precision and  $R$  is recall, where we define  $P$  to be  $Similarity(DG_1, DG_2)$  and  $R$  as  $Similarity(DG_2, DG_1)$ . The F-score is then used as a direct measurement of the similarity between two documents and in ranking the performance of different summarizers. Currently, we weight all the concepts/entities and relations equally.

### 3.2 Two groups of Document Sets

We generate a  $DG$  for each document in the DUC 2002 collection, and then calculate the value of similarity (F-score) between each pair of  $DG$ s within the same document set, and finally obtain the average similarity score. Using our similarity, we can naturally and automatically divide the document sets into two groups: similar document and diverse document sets. The similarity data demonstrated that most of the document sets in *one event, disaster domain, limited time* and *one event, any domain, limited time* categories are of the group of similar document sets, and most of the document sets in *multiple events, unlimited time* and *individual biography* categories are of diverse document group (as shown in Table 1).

Intuitively, these documents in *one event, disaster domain, limited time* and *one event, any domain, limited time* categories have more content overlap with each other and are more similar to each other than those in the other two categories.

**Table 1. The average F-scores for the document sets in different categories**

Category	One event disaster domain limited time	One event any domain limited time	Multiple events unlimited time	Individual biography
F-scores	0.10	0.12	0.06	0.07

An example of the group of similar document sets is *D.79.E.200.A* in this collection, which has been classified by DUC as *one event, disaster domain, limited time* category. It contains 9 articles, which are news reports about Hurricane Gilbert: when and where did it happen, and what kind of damage it caused. Obviously, the content of these news articles are highly overlapped. When creating the model

extracts for this set, one human assessor picked 7 sentences and another picked 8; between them, 4 sentences are actually exactly the same. The group of diverse document sets covers much broader topics. For example, the set *D.106.E.200.G* has been classified by DUC as *individual biography* category. It contains 10 articles consisting of stories such as: Northeastern University planning to award an honorary degree to Nelson Mandela; imprisoned African National Congress Leaders being allowed to visit Mandela; a family group visited Mandela; the Jewish-American group concerned about Mandela's attitude towards Israel and Palestine Liberation Organization; and, news that he would to be released.

The following are the two model extracts created for *D.106.E.200.G* set by human judges.

Assessor A: ONE MAN'S STRUGGLE

*From behind bars, Nelson Mandela has dominated the fight for black rights in South Africa. The following are some of the key events in his life.*

*Mandela joins the African National Congress at age 26, later becoming president of the group's Youth League. When the National Party comes to power in 1948, the ANC begins planning a campaign of civil disobedience to fight the party's apartheid policy. April-June, 1964 Mandela and seven others are sentenced to life in prison. December, 1989 President Frederik W. de Klerk meets Mandela for the first time. World leaders welcomed South Africa's announcement Saturday of Nelson R. Mandela's impending release from prison, and international rejoicing began to build for an event so long awaited by so many. Soweto and other black townships around Johannesburg have been hit by black factional fighting that has killed about 800 people since August. In the 15 months since he walked free from prison, Nelson Mandela has played the chief role in talks with President F.W. de Klerk and the white government. The radicals in the ANC may well use the trial and conviction of Mrs. Mandela to arouse members against Nelson Mandela's policy of peace and negotiation.*

*Assessor F: Mandela the 70-year-old leader of the ANC jailed 27 years ago, was sentenced to life in prison for conspiring to overthrow the South African government. Mandela was allowed to meet last Friday with five other ANC members convicted with him in 1964 of trying to overthrow the government and still imprisoned in Cape Town, according to Dullah Omar, a family friend and attorney. President Pieter W. Botha offers to free Mandela if he renounces violence; Mandela refuses until the government takes steps to dismantle apartheid and grants full political rights to blacks. De Klerk lifts ANC ban and says Mandela will be released. Commonwealth Secretary General Shridath Ramphal said Mandela's safety must be guaranteed by the South African government and that until change there becomes irreversible, the Commonwealth will sustain its pressure for the end of apartheid. The arrival Wednesday of African National Congress leader Nelson Mandela prompted an outpouring of praise and encouragement from politicians and supporters. In the 15 months since he walked free from prison, Nelson Mandela has played the chief role in talks with President F.W. de Klerk and the white government. Mandela and other ANC leaders have been negotiating with President F.W. de Klerk to pave the way for full-scale talks on ending apartheid, the South African system of racial separation.*

Assessor A spent the first half of the summary listing important events in the life of Nelson Mandela, then continued with his role in talks with the white government in

South Africa and the problems he was facing. While in Assessor F's summary, the information was more about what was happening at the time of the news concerning Mandela's release, after very briefly mentioning that he is an African National Congress (ANC) leader and has been jailed for 27 years. From the example, it is apparent that when facing such a complex task, people will try to pick up and organize information around their own biases. Just as central is what the judges ignored in their write-ups.

It would be very difficult for an automated summarizer to decide what should be covered in the summary. For the set *D.106.E.200.G*, two machine generated extracts are as follows:

*System 19: 1944 Mandela joins the African National Congress at age 26, later becoming president of the group's Youth League. When the National Party comes to power in 1948, the ANC begins planning a campaign of civil disobedience to fight the party's apartheid policy.*

*January, 1962.*

*In August he is charged with incitement and leaving the country illegally; gets five-year jail term.*

*July, 1963*

*Mandela, seven others are charged with sabotage and conspiracy to overthrow the government .*

*April-June, 1964*

*Mandela and seven others are sentenced to life in prison .*

*February, 1990*

*De Klerk lifts ANC ban and says Mandela will be released.*

*On Feb. 11, the nation's best-known black leader becomes a free man at last .*

*World leaders welcomed South Africa's announcement Saturday of Nelson R.*

*Mandela's impending release from prison, and international rejoicing began to build for an event so long awaited by so many.*

*Mandela has embraced Yasser Arafat, chairman of the Palestine Liberation Organization, in public and has compared the struggle of Palestinians to that of black South Africans.*

*Mayor David Dinkins: ``Our honored guest ... is a George Washington, a Bolivar, a King, a Herzl.*

*System 21: A 3-year campaign that has succeeded in getting Northeastern University to award an honorary degree to jailed South African nationalist leader Nelson Mandela has raised objections from some faculty and students.*

*A representative of the African National Congress said Saturday the South African government may release black nationalist leader Nelson Mandela as early as Tuesday.*

*Four imprisoned African National Congress leaders, convicted in 1964 with Nelson Mandela, were allowed to visit their ANC colleague at his isolated prison house, an activist said today.*

*The largest family group to visit black leader Nelson Mandela since he was jailed 27 years ago gathered in Cape Town for a meeting Tuesday with the prisoner on his 71st birthday.*

*World leaders welcomed South Africa's announcement Saturday of Nelson R.*

*Mandela's impending release from prison, and international rejoicing began to build for an event so long awaited by so many.*

*Mayor David Dinkins warned Jewish groups against protesting the visit of anti-apartheid leader Nelson Mandela to New York City, saying the protests might insult the black community, a newspaper said today.*

*Unidentified assailants hurled a grenade and fired shots today at the home of relatives of Nelson Mandela, killing a baby girl and injuring her parents, police said.*

For System 19, it tried to include the important events in Mandela's life and his fight against apartheid. While, for System 21, it is more likely that it picked one sentence from every article in the document set without any preference or focus.

In general, the group of similar document sets consists of highly correlated documents forming a more coherent focus; and the group of diverse document sets contains more diverse documents covering a wide scope of topics. This would naturally suggest that, it would be more difficult to generate quality summaries for the group of diverse document sets, since summaries would need to cover more topics. Our experiments in the next section demonstrate otherwise.

## **4. Comparing rankings obtained from different approaches for evaluation on the impact of document sets.**

### **4.1 Hypothesis**

As mentioned above, there are two general groups of document sets in the DUC 2002 collection, which suggests that obtaining good summaries for the group of diverse document sets is more difficult than for those documents belonging to the group of similar document sets. If we assume that some automated summarizers can generate quality summaries for both groups, while others performed worse for one group (more likely group of diverse document sets); then the ranking orders of summarizer performance by human assessors should be different, when they work with document sets that belong to different groups. There is a possibility that all summarizers performed similarly good or bad when working with certain document sets, however, from our examination of the automatically generated summaries, we believe this possibility to be low.

### **4.2 Experimental Procedure**

Three different approaches have been used to rank the performance of summarizers.

(i) Exact sentence matching, where system generated extracts were compared to model summaries created by human (which are also included in DUC 2002 data).

(ii) Document graph comparison, where automatically generated extracts are compared with the original documents based on the document graphs that generated, and average F-scores were calculated for comparison and ranking purposes [19].

(iii) Human judgments on the quality of the summaries, other than just comparing machine generated summaries with human created model summaries. It was expected that human evaluators would clearly recognize the difference between the two types of document sets.

Five people participated in the experiment. They are all graduate students pursuing their PhDs in computer science. One of them is a native English speaker, while the other four participants are from Asia and North Africa. Although English is a second language for four participants, they have no problem to understand general news reports in English since they have been living in the U.S. and studying towards advanced science degrees for at least 5 five years.

Each participant was given 4 document sets, which were randomly picked from the 59 document sets in the 2002 DUC data collection; together with extracts generated by 10 automated summarizers for each document set. The summarizers have been numbered by DUC as systems 16, 19, 20, 21, 22, 24, 25, 28, 29, and 31 (Table 2). To avoid possible bias, the extracts have been renumbered. For example, in document set 61, summary number 1 is generated by System 22, but in document set 62, summary number 1 is generated by System 20. In total, 20 document sets have been evaluated in the experiment, within which 11 are belong to the group of similar document sets and 9 are belong to the group of diverse document sets.

The participants also received instruction on how to evaluate the extracts. They were asked to read carefully through the original document sets and identify the information they think that should be covered in a summary. After reviewing extracts, they are asked to assign a score to each of the extracts using a five-point-scale scoring system, in which from 1 to 5, the quality of the summary would be from very poor to very good, respectively. They were instructed that for a good summary, its quality is based on how well it covers important information, while the order of the sentences and the transition between two sentences in the extracts are not important factors. The time spent on each document set was 53.2 minutes on average (ranging from 31 minutes to 257 minutes, depending on the length the documents).

The rankings obtained based on the different approaches have been compared by using the Spearman rank correlation coefficient ( $r$ ) [6].

### 4.3 Results

Note that three different ranking approaches have been applied and compared in our experiments. Sentence matching compares the machine generated summaries and the model summaries created by human; *DG* approach measures the information coverage of the extracts; and human evaluation is based on direct human judgment.

The ranking results are shown in Table 2, and the correlation data for different ranking approaches are shown in Table 3. Different ranking approaches ended up with different results, only the ranking from sentence matching and human judgment for the group of similar document sets demonstrated that they are highly correlated.

For the group of similar document sets, sentence matching and human judgment gave very similar ranking results ( $r = 0.92$ ,  $p < 0.01$ ), which were different from the *DG* approach results. This demonstrates to us the difference among the three ranking approaches. Both sentence matching and human judgment involve human opinion on what is important, while the *DG* approach simply measures the information coverage. However, when applied to the group of diverse document sets, the correlation between sentence matching and human judgment was no longer statistically significant (Table 3), which may due to the reason that summarization for the group of diverse document sets is a more complex task and, hence, it is more difficult to

reach agreement on which summarizers performed better from different ranking approaches.

**Table 2. Ranking order obtained based on three different approaches (*DG*: document graph comparison, *S*: sentence matching, *H*: Human judger scoring).**

System	Group of similar document sets			Group of diverse document sets		
	<i>DG</i>	<i>S</i>	<i>H</i>	<i>DG</i>	<i>S</i>	<i>H</i>
16	8	9	9	6	8	10
19	3	1	2	4	6	3.5
20	6	4	4	5	4	5
21	4	3	3	3	1	1.5
22	9	10	10	10	10	9
24	7	2	1	1	3	3.5
25	5	7	7.5	8	9	6
28	10	5	7.5	9	2	8
29	2	6	5	2	5	7
31	1	8	6	7	7	1.5

We assumed that some summarizers could generate quality summaries for both groups of document sets while others could have a fair performance for the group of similar document sets but not for the group of diverse document sets. Thus, the ranking orders for the summarizers would be different when evaluated based on different types of document sets, especially by human judgment. The experiments showed that the *DG* approach indicated that there is a significant difference in summarizer performance when working on the group of similar document sets versus the group of diverse document sets. The Spearman rank correlation coefficient of the ranking orders obtained when working with both groups is only 0.45, and the correlation is not statistically significant (Table 4). In other words, at least some summarizers performed differently with two different groups of document sets. However, the sentence matching approach could not identify this difference as clearly as the *DG* approach, the correlation coefficient for the two rankings being 0.72, and the *p-value* indicted the correlation is statistically significant. Surprisingly, the human also failed to realize the difference ( $r = 0.75$  and  $p = 0.006$ ) (Table 4).

Although we can not completely rule out the possibility that all 10 automatic summarizers performed similarly well with the group of similar document sets but not the group of diverse document sets, the low correlation between the two ranking orders for different types of document sets based on *DG* still supported our hypothesis.

**Table 3. Correlation between the rankings obtained based on three different approaches (\*  $p < 0.05$ , \*\*  $p < 0.01$ ).**

		<i>DG</i>	<i>S</i>	<i>H</i>
Group of similar documents	<i>DG</i>	-		
	<i>S</i>	0.21	-	
	<i>H</i>	0.45	0.92**	-
Group of diverse documents	<i>DG</i>	-		
	<i>S</i>	0.54	-	
	<i>H</i>	0.48	0.44	-

**Table 4. Correlation between the rankings obtained by the same method on different groups of document sets.**

<i>DG</i>	<i>S</i>	<i>H</i>
0.45	0.72**	0.75**

Previous studies showed that human judgment differences could be one of the variations that affect the performance scores [3]. In our experiments, five participants generally agreed with each other on the performance of the automatic summarizers, where in pair-wised comparisons, 7 out of 10 pairs showed the correlation between two human judges was statistically significant (Table 5). Also, in this experiment, not only were the extracts provided by the 10 summarizers included, but also included the model summaries. In fact, the model summaries generated by humans were always ranked as the best ones, except in one case where it ranked as the third best; which indicated that human participants were doing a good job on evaluating the summary quality.

**Table 5. Pair-wised correlation analysis on ranking orders obtained based on the scores assigned from different judges (The experiments were performed on all data, including both groups of document sets) (\*  $p < 0.05$ , \*\*  $p < 0.01$ ).**

Judge	1	2	3	4	5
1	-				
2	0.37	-			
3	0.80**	0.39	-		
4	0.60*	0.62*	0.64*	-	
5	0.92**	0.45	0.67*	0.57*	-

In our user evaluation experiments, in trying to eliminate possible bias, we told the participants that a quality summary should cover important content, but deliberately avoided directing what kind of content should be considered as important (is it high level analysis? or is it detailed information? or where the focus should be? etc.). Unfortunately, this also resulted in the fact that no specific task was assigned for the summarization process. In this situation, human judges might assign the same scores

to the summaries that covered some general topics and detailed information, although their coverage and focuses could be quite different. In evaluations using model summaries, it is a common practice to use summaries created by several judges, such as in DUC 2002 data collection. Here the model summaries have been created by 9 judges with each document set having 2 model summaries. Then, the individual differences among the judges may be masked by the compensation from different people.

Furthermore, we calculated pair-wised correlations among five human judges separately on the group of similar document set and the group of diverse document sets. When working with the group of diverse document sets, there is only 1 pair in a total of 10 that correlated with each other at a level of statistically significant, with average  $r = 0.19$ . For the group of similar document sets, there are 4 pairs rankings showing statistically significant correlation, and the average  $r$  is increased to 0.47. It suggested that, as individuals, the human judges had more disagreements with each other when working with diverse document sets.

## 5. Discussion and conclusion

To participate in preventing or responding to a terrorism task, intelligent analysts often face a real challenge in finding critical information and making decision under time pressure. Multi-document summarization is a very useful technique to assist intelligence analysts in their daily work to find relevant information from the massive amount of available data. The process of automatic summarization can be decomposed into three steps: analyzing the input text, extracting important information, and synthesizing an appropriate output. As we talked earlier, a summarization should be user- and task-oriented. Therefore, understanding human and their needs are crucial for a good text summarization system.

Various technologies have been attempted for generating summaries, such as term frequency [4], predefined templates [8], and, latent semantic indexing [2]. However, much less has been done in identifying how individual differences affect the perception of a good/quality summarization. There has been a closely related effort that used the utility of query biased summaries to help users identify relevant documents [14], in which Local Context Analysis (LCA) has been used to expand topics contained in the baseline summaries with additional words and phrases. LCA is a technique for automatic query expansion using pseudo feedback. It examines the context surrounding the topic terms in the top ranked documents for query expansion. It has been found that users could judge the relevance of documents based on their summaries, almost as accurately as if they accessed the full texts. Sakai and Masuyama [12] proposed an interactive approach for multi-document summarization realizing a user's summarization need. Their system extracts keywords from a document set and shows  $k$  best keywords with scores to a user on the screen. Then the user has the opportunity to select those that reflect his/her information needs. The approach helped improve the system performance.

Our *DG* approach can automatically identify if a summary is created from a broad and diverse document sets (as opposed to a highly focused set). It then can serve as an alert to the user when there is a high risk of missing information the user may be interested in. In our experiments, human evaluators, as a group, were not very

sensitive to this difference; however, as individuals, they had more disagreement with each other when they working with document sets covering diverse content. This result, again, suggested that each individual has his/her own information needs. We also believe that the difference in individual information needs would be much more obvious when they working on certain tasks or have specific goals.

More importantly, the difference between these two groups of document sets and the human response to it actually reflects the requirements for a good summary. People generally agree with each other on the most common content needed to be covered in a summary; and, they do have their own individual interests. Therefore, a quality summary should cover the most general topics and also various related pieces of information from different aspects or details that are relevant to the user's specific needs. Thus, for user-centered summarization, an important task is to identify what kind of the details in information needs to be retrieved and kept in the summary to satisfy the specific user interests according to the user's knowledge and the specific tasks he/she has.

McKeown et al [7] found that for the multi-event input document sets, the difficulty in generating a quality summary comes from the breadth and diversity of the documents in the sets. Sometimes, even humans seem to have a hard time determining how to produce a good summary. In this situation, their summaries were often quite different from each other (as showed by the example in Section 3.2). The method proposed to handle this situation is to apply different strategies/techniques on different categories of documents in the input sets. For DUC 2002 data, they used four different strategies, one for single events, one for multiple related events, one for biographies, and one for discussion of an issue with related events [7]. The problem associated with this approach is that, the categories are classified based on the content of the documents. There should be more categories than the four being used in DUC 2002 collection, and a method is also needed to automatically classify documents. McKeown et al. [7] suggest create document sets and define a set of criteria by automatically filtering and clustering large online data [7], while it is a manual process in DUC 2002. In addition, the user's needs have not been considered.

Our *DG* approach has the potential to overcome this problem. *DG* generation is a process of information extraction and representation. As a result, the important concepts, entities and relations in the text are captured. We can then generate a summary from each *DG* of each document. First, a core of a summary (also in *DG* format) that contains the most general information is needed. For document sets, the core can be constructed by majority vote; while for a single document, the core can be created based on the weights of the relations. We then expand the core *DG* by inserting relevant relations based on the underlying graph structure. The most relevant relations would be decided with the help from a user model that captures a user's knowledge and foci/interests, which answers to the challenge that a good summary should target who reads it. Finally, we generate a summary based on this *DG*, which should be biased towards a user's individual interests, and better meet his/her information needs. This approach naturally fits into our previous efforts on using a user model to provide proactive assistance in information searching process [15]. We are currently pursuing this effort and are focused on formally defining the appropriate graph theoretic measures for expanding *DGs* from multiple documents.

## 6. ACKNOWLEDGMENTS

This work was supported in part by the Advanced Research and Development Activity (ARDA) U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government.

## 7. REFERENCES

- [1] Elhadad, N., "User-Sensitive Text Summarization", AAAI Doctoral Consortium, San Jose, CA. pp. 987-988, 2004
- [2] Gong, Y. and Liu, X., "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis". In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp 19-25. New Orleans, Louisiana. 2001.
- [3] Harman, D., and Over, P., "The effects of human variation in duc summarization evaluation". In Proceedings of ACL 2004, Workshop on Text Summarization Branches out, 2004.
- [4] Luhn, H.P. "The automatic creation of literature abstracts". IBM Journal, pp 159-165, 1958
- [5] Mani, I. and Maybury, M.T., *Advances in Automatic Text Summarization*. The MIT Press. 1999
- [6] Myers, J. L. and Well, A. D., *Research Design and Statistical Analysis*. Lawrence Erlbaum Associates, publishers, New Jersey pp. 488-490, 1995
- [7] McKeown, K., Barzilay, R. and Blair-Goldensohn, S., "The Columbia Multi-Document Summarizer for DUC 2002". 2002
- [8] McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C, Schiffman, B. and Sigelman, S., "Tracking and Summarizing News on Daily Basis with Columbia's Newsblaster." In Proceedings of 2002 Human Language Technology Conference (HLT), San Diego, CA, 2002
- [9] McKeown, K., Passonneau R.J. and Elson D.K., "Do Summaries Help? A Task-Based Evaluation of Multi-Document Summarization." SIGIR'05, Salvador, Brazil, 2005
- [10] McKeown, K. and Radev, D.R., "Generating Summaries from Multiple News Articles". In Proceedings of the 18th Annual ACM SIGIR Conference, pp. 74-82. Seattle, WA. 1995
- [11] Montes-y-Gómez, M., Gelbukh, A., and López-López, A., "Comparison of Conceptual Graphs." In Proceeding of MICAI-2000 – 1st Mexican International Conference on Artificial Intelligence. Acapulco, Mexico, 2000
- [12] Over, P. and Liggett, W., "Introduction to DUC-2002: an Intrinsic Evaluation of Generic News Text Summarization Systems". Document Understanding Conference website (<http://duc.nist.gov/>), 2002
- [13] Sakai, H. and Masuyama, S., "A Multiple-Document Summarization System Introducing User Interaction for reflecting User's Summarization Need". Working Notes of NTCIR-4, Tokyo, 2-4 June 2004.

- [14] Sanderson, M., "Accurate User Directed Summarization from Existing Tools". In Proceedings of the 7th International Conference on Information and Knowledge Management. Bethesda, Maryland. pp 45-51, 1998
- [15] Santos, E.Jr., Nguyen, H., Zhao, Q and Wang, H, "User modelling for intent prediction in information analysis. In Proceedings of the 47th Annual Meeting for the Human Factors and Ergonomics Society, pp 1034–1038, 2003.
- [16] Santos, E. Jr., Nguyen, H., Zhao, Q. and Pukinskis, E. Empirical Evaluation of Adaptive User Modeling in a Medical Information Retrieval Application. Lecture Notes in Artificial Intelligence 2702: User Modeling 2003 (Eds. P. Brusilovsky, A. Corbett, and F. de Rosis), Springer. Pages 292-296. 2003
- [17] Santos, E. Jr., Mohamed, A.A and Zhao, Q., "Automatic Evaluation of Summaries Using Document Graphs". In Proceedings of ACL 2004, Workshop on Text Summarization Branches out, Barcelona, Spain, pp 66-73, 2004
- [18] Sleator, D. D. and Temperley, D., "Parsing English with a link grammar". In Proceedings of the 3rd International Workshop on Parsing Technologies. pp 277-292, 1993
- [19] Van Rijsbergen Keith. Information Retrieval. 2nd Edition. Butterworths, London, 1979