Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization

Ani Nenkova
University of Pennsylvania
Philadelphia, PA 19104, USA
nenkova@seas.upenn.edu

Annie Louis
University of Pennsylvania
Philadelphia, PA 19104, USA
lannie@seas.upenn.edu

Abstract

Different summarization requirements could make the writing of a good summary more difficult, or easier. Summary length and the characteristics of the input are such constraints influencing the quality of a potential summary. In this paper we report the results of a quantitative analysis on data from large-scale evaluations of multi-document summarization, empirically confirming this hypothesis. We further show that features measuring the cohesiveness of the input are highly correlated with eventual summary quality and that it is possible to use these as features to predict the difficulty of new, unseen, summarization inputs.

1 Introduction

In certain situations even the best automatic summarizers or professional writers can find it hard to write a good summary of a set of articles. If there is no clear topic shared across the input articles, or if they follow the development of the same event in time for a longer period, it could become difficult to decide what information is most representative and should be conveyed in a summary. Similarly, length requirements could pre-determine summary quality—a short outline of a story might be confusing and unclear but a page long discussion might give an excellent overview of the same issue.

Even systems that perform well on average produce summaries of poor quality for some inputs. For this reason, understanding what aspects of the input make it difficult for summarization becomes an interesting and important issue that has not been addressed in the summarization community untill now. In information retrieval, for example, the variable system performance has been recognized as a research challenge and numerous studies on identifying query difficulty have been carried out (most recently (Cronen-Townsend et al., 2002; Yom-Tov et al., 2005; Carmel et al., 2006)).

In this paper we present results supporting the hypotheses that input topicality cohesiveness and summary length are among the factors that determine summary quality regardless of the choice of summarization strategy (Section 2). The data used for the analyses comes from the annual Document Understanding Conference (DUC) in which various summarization approaches are evaluated on common data, with new test sets provided each year.

In later sections we define a suite of features capturing aspects of the topicality cohesiveness of the input (Section 3) and relate these to system performance, identifying reliable correlates of input difficulty (Section 4). Finally, in Section 5, we demonstrate that the features can be used to build a classifier predicting summarization input difficulty with accuracy considerably above chance level.

2 Preliminary analysis and distinctions: DUC 2001

Generic multi-document summarization was featured as a task at the Document Understanding Conference (DUC) in four years, 2001 through 2004. In our study we use the DUC 2001 multi-document task submissions as development data for in-depth analysis and feature selection. There were 29 input sets and 12 automatic summarizers participating in the evaluation that year. Summaries of different

Proceedings of ACL-08: HLT, pages 825–833,
Columbus, Ohio, USA, June 2008.

lengths were produced by each system: 50, 100, 200 and 400 words. Each summary was manually evaluated to determine the extent to which its content overlaped with that of a human model, giving a *coverage score*. The content comparison was performed on a subsentence level and was based on elementary discourse units in the model summary.[1]

The coverage scores are taken as an indicator of difficultly of the input: systems achieve low coverage for difficult sets and higher coverage for easy sets. Since we are interested in identifying characteristics of generally difficult inputs rather than in discovering what types of inputs might be difficult for one given system, we use the average system score per set as indicator of general difficulty.

2.1 Analysis of variance

Before attempting to derive characteristics of inputs difficult for summarization, we first confirm that indeed expected performance is influenced by the input itself. We performed analysis of variance for DUC 2001 data, with *automatic system* coverage score as the dependent variable, to gain some insight into the factors related to summarization difficulty. The results of the ANOVA with input set, summarizer identity and summary length as factors, as well as the interaction between these, are shown in Table 1.

As expected, summarizer identity is a significant factor: some summarization strategies/systems are more effective than others and produce summaries with higher coverage score. More interestingly, the input set and summary length factors are also highly significant and explain more of the variability in coverage scores than summarizer identity does, as indicated by the larger values of the F statistic.

Length The average automatic summarizer coverage scores increase steadily as length requirements are relaxed, going up from 0.50 for 50-word summaries to 0.76 for 400-word summaries as shown in Table 2 (second row). The general trend we observe is that on average systems are better at producing summaries when more space is available. The differences are statistically significant[2] only between 50-word and 200- and 400-word summaries and between 100-word and 400-word summaries. The fact that summary quality improves with increasing summary length has been observed in prior studies as well (Radev and Tam, 2003; Lin and Hovy, 2003b; Kolluru and Gotoh, 2005) but generally little attention has been paid to this fact in system development and no specific user studies are available to show what summary length might be most suitable for specific applications. In later editions of the DUC conference, only summaries of 100 words were produced, focusing development efforts on one of the more demanding length restrictions. The interaction between summary length and summarizer is small but significant (Table 1), with certain summarization strategies more successful at particular summary lengths than at others.

Type	**50**	**100**	**200**	**400**
Human	1.00	1.17	1.38	1.29
Automatic	0.50	0.55	0.70	0.76
Baseline	0.41	0.46	0.52	0.57

Table 2: Average human, system and baseline coverage scores for different summary lengths of N words. $N =$ 50, 100, 200, and 400.

Improved performance as measured by increase in coverage scores is observed for human summarizers as well (shown in the first row of Table 2). Even the baseline systems (first n words of the most recent article in the input or first sentences from different input articles) show improvement when longer summaries are allowed (performance shown in the third row of the table). It is important to notice that the difference between automatic system and baseline performance increases as the summary length increases—the difference between systems and baselines coverage scores is around 0.1 for the shorter 50- and 100-word summaries but 0.2 for the longer summaries. This fact has favorable implications for practical system developments because it indicates that in applications where somewhat longer summaries are appropriate, automatically produced summaries will be much more informative than a baseline summary.

[1] The routinely used tool for automatic evaluation ROUGE was adopted exactly because it was demonstrated it is highly correlated with the manual DUC coverage scores (Lin and Hovy, 2003a; Lin, 2004).

[2] One-sided t-test, 95% level of significance.

Factor	DF	Sum of squares	Expected mean squares	F stat	Pr($> F$)
input	28	150.702	5.382	59.4227	0
summarizer	11	34.316	3.120	34.4429	0
length	3	16.082	5.361	59.1852	0
input:summarizer	306	65.492	0.214	2.3630	0
input:length	84	36.276	0.432	4.7680	0
summarizer:length	33	6.810	0.206	2.2784	0

Table 1: Analysis of variance for coverage scores of automatic systems with input, summarizer, and length as factors.

Input The input set itself is a highly significant factor that influences the coverage scores that systems obtain: some inputs are handled by the systems better than others. Moreover, the input interacts both with the summarizers and the summary length.

This is an important finding for several reasons. First, in system evaluations such as DUC the inputs for summarization are manually selected by annotators. There is no specific attempt to ensure that the inputs across different years have on average the same difficulty. Simply assuming this to be the case could be misleading: it is possible in a given year to have "easier" input test set compared to a previous year. Then system performance across years cannot be meaningfully compared, and higher system scores would not be indicative of system improvement between the evaluations.

Second, in summarization applications there is some control over the input for summarization. For example, related documents that need to summarized could be split into smaller subsets that are more amenable to summarization or routed to an appropriate summarization system than can handle this kind of input using a different strategy, as done for instance in (McKeown et al., 2002).

Because of these important implications we investigate input characteristics and define various features distinguishing easy inputs from difficult ones.

2.2 Difficulty for people and machines

Before proceeding to the analysis of input difficulty in multi-document summarization, it is worth mentioning that our study is primarily motivated by system development needs and consequently the focus is on finding out what inputs are easy or difficult *for automatic systems*. Different factors might make summarization difficult *for people*. In order to see to what extent the notion of summarization input difficulty is shared between machines and people, we computed the correlation between the average system and average human coverage score at a given summary length for all DUC 2001 test sets (shown in Table 3). The correlation is highest for 200-word summaries, 0.77, which is also highly significant. For shorter summaries the correlation between human and system performance is not significant.

summary length	correlation
50	0.50
100	0.57*
200	0.77**
400	0.70**

Table 3: Pearson correlation between average human and system coverage scores on the DUC 2001 dataset. Significance levels: *$p < 0.05$ and **$p < 0.00001$.

In the remaining part of the paper we deal exclusively with difficulty as defined by system performance, which differs from difficulty for people summarizing the same material as evidenced by the correlations in Table 3. We do not attempt to draw conclusions about any cognitively relevant factors involved in summarizing.

2.3 Type of summary and difficulty

In DUC 2001, annotators prepared test sets from five possible predefined input categories:[3].

Single event (3 sets) Documents describing a single event over a timeline (e.g. The Exxon Valdez oil spill).

[3] Participants in the evaluation were aware of the different categories of input and indeed some groups developed systems that handled different types of input employing different strategies (McKeown et al., 2001). In later years, the idea of multi-strategy summarization has been further explored by (Lacatusu et al., 2006)

Subject (6 sets) Documents discussing a single topic (e.g. Mad cow disease)

Biographical (2 sets) All documents in the input provide information about the same person (e.g. Elizabeth Taylor)

Multiple distinct events (12 sets) The documents discuss different events of the same type (e.g. different occasions of police misconduct).

Opinion (6 sets) Each document describes a different perspective to a common topic (e.g. views of the senate, congress, public, lawyers etc on the decision by the senate to count illegal aliens in the 1990 census).

Figure 1 shows the average system coverage score for the different input types. The more topically cohesive input types such as *biographical*, *single event* and *subject*, which are more focused on a single entity or news item and narrower in scope, are easier for systems. The average system coverage score for them is higher than for the non-cohesive sets such as multiple distinct events and opinion sets, regardless of summary length. The difference is even more apparently clear when the scores are plotted after grouping input types into cohesive (biographical, single event and subject) and non-cohesive (multiple events and opinion). Such grouping also gives the necessary power to perform statistical test for significance, confirming the difference in coverage scores for the two groups. This is not surprising: a summary of documents describing multiple distinct events of the same type is likely to require higher degree of generalization and abstraction. Summarizing opinions would in addition be highly subjective. A summary of a cohesive set meanwhile would contain facts directly from the input and it would be easier to determine which information is important. The example human summaries for set D32 (single event) and set D19 (opinions) shown below give an idea of the potential difficulties automatic summarizers have to deal with. **set D32** On 24 March 1989, the oil tanker Exxon Valdez ran aground on a reef near Valdez, Alaska, spilling 8.4 million gallons of crude oil into Prince William Sound. In two days, the oil spread over 100 miles with a heavy toll on wildlife. Cleanup proceeded at a slow pace, and a plan for cleaning 364 miles of Alaskan coastline was released. In June, the tanker was refloated. By early 1990, only 5 to 9 percent of spilled oil was recovered. A federal jury indicted Exxon on five criminal charges and the Valdez skipper was guilty of negligent discharge of oil.

set D19 Congress is debating whether or not to count illegal aliens in the 1990 census. Congressional House seats are apportioned to the states and huge sums of federal money are allocated based on census population. California, with an estimated half of all illegal aliens, will be greatly affected. *Those arguing for inclusion say that the Constitution does not mention "citizens", but rather, instructs that House apportionment be based on the "whole number of persons" residing in the various states. Those opposed say that the framers were unaware of this issue. "Illegal aliens" did not exist in the U.S. until restrictive immigration laws were passed in 1875.*

The manual set-type labels give an intuitive idea of what factors might be at play but it is desirable to devise more specific measures to predict difficulty. Do such measures exist? Is there a way to automatically distinguish cohesive (easy) from non-cohesive (difficult) sets? In the next section we define a number of features that aim to capture the cohesiveness of an input set and show that some of them are indeed significantly related to set difficulty.

3 Features

We implemented 14 features for our analysis of input set difficulty. The working hypothesis is that cohesive sets with clear topics are easier to summarize and the features we define are designed to capture aspects of input cohesiveness.

Number of sentences in the input, calculated over all articles in the input set. Shorter inputs should be easier as there will be less information loss between the summary and the original material.

Vocabulary size of the input set, equal to the number of unique words in the input. Smaller vocabularies would be characteristic of easier sets.

Percentage of words used only once in the input. The rationale behind this feature is that cohesive input sets contain news articles dealing with a clearly defined topic, so words will be reused across documents. Sets that cover disparate events and opinions are likely to contain more words that appear in the input only once.

Type-token ratio is a measure of the lexical variation in an input set and is equal to the input vocabulary size divided by the number of words in the

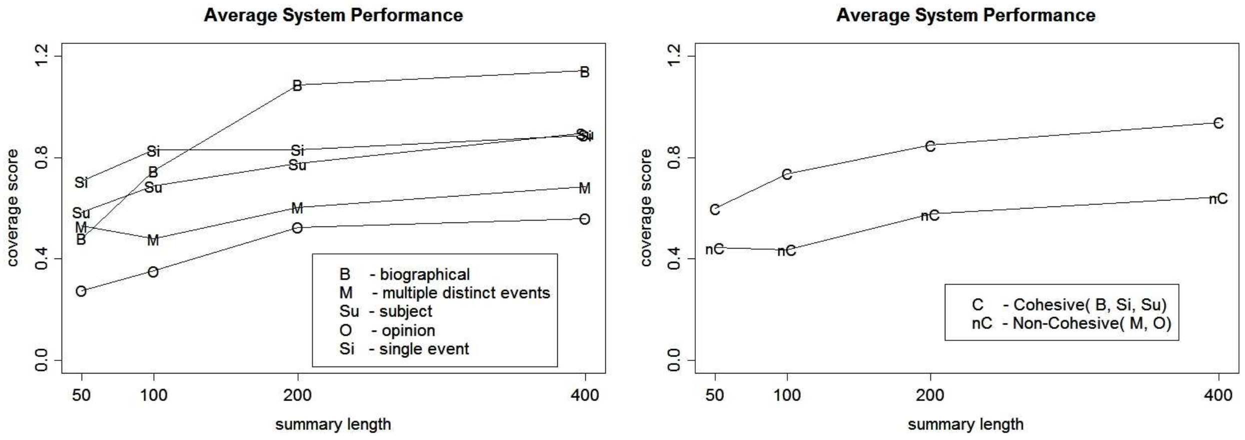

Figure 1: Average system coverage scores for summaries in a category

input. A high type-token ratio indicates there is little (lexical) repetition in the input, a possible side-effect of non-cohesiveness.

Entropy of the input set. Let X be a discrete random variable taking values from the finite set $V = \{w_1, ..., w_n\}$ where V is the vocabulary of the input set and w_i are the words that appear in the input. The probability distribution $p(w) = Pr(X = w)$ can be easily calculated using frequency counts from the input. The entropy of the input set is equal to the entropy of X:

$$H(X) = -\sum_{i=1}^{i=n} p(w_i) \log_2 p(w_i) \qquad (1)$$

Average, minimum and maximum cosine overlap between the news articles in the input. Repetition in the input is often exploited as an indicator of importance by different summarization approaches (Luhn, 1958; Barzilay et al., 1999; Radev et al., 2004; Nenkova et al., 2006). The more similar the different documents in the input are to each other, the more likely there is repetition across documents at various granularities.

Cosine similarity between the document vector representations is probably the easiest and most commonly used among the various similarity measures. We use tf*idf weights in the vector representations, with term frequency (tf) normalized by the total number of words in the document in order to remove bias resulting from high frequencies by virtue of higher document length alone.

The cosine similarity between two (document representation) vectors v_1 and v_2 is given by $cos\theta = \frac{v_1.v2}{||v_1||||v_2||}$. A value of 0 indicates that the vectors are orthogonal and dissimilar, a value of 1 indicates perfectly similar documents in terms of the words contained in them.

To compute the cosine overlap features, we find the pairwise cosine similarity between each two documents in an input set and compute their average. The minimum and maximum overlap features are also computed as an indication of the overlap bounds. We expect cohesive inputs to be composed of similar documents, hence the cosine overlaps in these sets of documents must be higher than those in non-cohesive inputs.

KL divergence Another measure of relatedness of the documents comprising an input set is the difference in word distributions in the input compared to the word distribution in a large collection of diverse texts. If the input is found to be largely different from a generic collection, it is plausible to assume that the input is not a random collection of articles but rather is defined by a clear topic discussed within and across the articles. It is reasonable to expect that the higher the divergence is, the easier it is to define what is important in the article and hence the easier it is to produce a good summary.

For computing the distribution of words in a general background corpus, we used all the inputs sets from DUC years 2001 to 2006. The divergence measure we used is the Kullback Leibler divergence, or relative entropy, between the input (I) and collection language models. Let $p_{inp}(w)$ be the probability of the word w in the input and $p_{coll}(w)$ be the probability of the word occurring in the large background collection. Then the relative entropy between the input and the collection is given by

$$\text{KL divergence} = \sum_{w \in I} p_{inp}(w) \log_2 \frac{p_{inp}(w)}{p_{coll}(w)} \quad (2)$$

Low KL divergence from a random background collection may be characteristic of highly non-cohesive inputs consisting of unrelated documents.

Number of topic signature terms for the input set. The idea of topic signature terms was introduced by Lin and Hovy (Lin and Hovy, 2000) in the context of single document summarization, and was later used in several multi-document summarization systems (Conroy et al., 2006; Lacatusu et al., 2004; Gupta et al., 2007).

Lin and Hovy's idea was to automatically identify words that are descriptive for a cluster of documents on the same topic, such as the input to a multi-document summarizer. We will call this cluster T. Since the goal is to find descriptive terms for the cluster, a comparison collection of documents not on the topic is also necessary (we will call this background collection NT).

Given T and NT, the likelihood ratio statistic (Dunning, 1994) is used to identify the topic signature terms. The probabilistic model of the data allows for statistical inference in order to decide which terms t are associated with T more strongly than with NT than one would expect by chance.

More specifically, there are two possibilities for the distribution of a term t: either it is very indicative of the topic of cluster T, and appears more often in T than in documents from NT, or the term t is not topical and appears with equal frequency across both T and NT. These two alternatives can be formally written as the following hypotheses:

H1: $P(t|T) = P(t|NT) = p$ (t is not a descriptive term for the input)

H2: $P(t|T) = p_1$ and $P(t|NT) = p_2$ and $p_1 > p_2$ (t is a descriptive term)

In order to compute the likelihood of each hypothesis given the collection of the background documents and the topic cluster, we view them as a sequence of words w_i: $w_1 w_2 \ldots w_N$. The occurrence of a given word t, $w_i = t$, can thus be viewed a Bernoulli trial with probability p of success, with success occurring when $w_i = t$ and failure otherwise.

The probability of observing the term t appearing k times in N trials is given by the binomial distribution

$$b(k, N, p) = \binom{N}{k} p^k (1-p)^{N-k} \quad (3)$$

We can now compute

$$\lambda = \frac{\text{Likelihood of the data given H1}}{\text{Likelihood of the data given H2}} \quad (4)$$

which is equal to

$$\lambda = \frac{b(c_t, N, p)}{b(c_T, N_T, p_1) * b(c_{NT}, N_{NT}, p_2)} \quad (5)$$

The maximum likelihood estimates for the probabilities can be computed directly. $p = \frac{c_t}{N}$, where c_t is equal to the number of times term t appeared in the entire corpus T+NT, and N is the number of words in the entire corpus. Similarly, $p_1 = \frac{c_T}{N_T}$, where c_T is the number of times term t occurred in T and N_T is the number of all words in T. $p_2 = \frac{c_{NT}}{N_{NT}}$, where c_{NT} is the number of times term t occurred in NT and N_{NT} is the total number of words in NT.

$-2log\lambda$ has a well-know distribution: χ^2. Bigger values of $-2log\lambda$ indicate that the likelihood of the data under H2 is higher, and the χ^2 distribution can be used to determine when it is significantly higher ($-2log\lambda$ exceeding 10 gives a significance level of 0.001 and is the cut-off we used).

For terms for which the computed $-2log\lambda$ is higher than 10, we can infer that they occur more often with the topic T than in a general corpus NT, and we can dub them "topic signature terms".

Percentage of signature terms in vocabulary The number of signature terms gives the total count of topic signatures over all the documents in the input. However, the number of documents in an input set and the size of the individual documents across different sets are not the same. It is therefore possible that the mere count feature is biased to the length and number of documents in the input set. To account for this, we add the percentage of topic words in the vocabulary as a feature.

Average, minimum and maximum topic signature overlap between the documents in the input. Cosine similarity measures the overlap between two documents based on all the words appearing in them. A more refined document representation can be defined by assuming the document vectors contain only the topic signature words rather than all words. A high overlap of topic words across two documents is indicative of shared topicality. The average, minimum and maximum pairwise cosine overlap between the tf*idf weighted topic signature vectors of the two documents are used as features for predicting input cohesiveness. If the overlap is large, then the topic is similar across the two documents and hence their combination will yield a cohesive input.

feature	**t-stat**	**p-value**
KL divergence*	-2.4725	*0.01*
% of sig. terms in vocab*	-2.0956	*0.02*
average cosine overlap*	-2.1227	*0.02*
vocabulary size*	1.9378	*0.03*
set entropy*	2.0288	*0.03*
average sig. term overlap*	-1.8803	*0.04*
max cosine overlap	-1.6968	0.05
max topic signature overlap	-1.6380	0.06
number of sentences	1.4780	0.08
min topic signature overlap	-0.9540	0.17
number of signature terms	0.8057	0.21
min cosine overlap	-0.2654	0.39
% of words used only once	0.2497	0.40
type-token ratio	0.2343	0.41

$*$Significant at a 95% confidence level($p < 0.05$)

Table 4: Comparison of non-cohesive (average system coverage score $<$ median average system score) vs cohesive sets for summary length of 100 words

4 Feature selection

Table 4 shows the results from a one-sided t-test comparing the values of the various features for the easy and difficult input set classes. The comparisons are for summary length of 100 words because in later years only such summaries were evaluated. The binary easy/difficult classes were assigned based on the average system coverage score for the given set, with half of the sets assigned to each class.

In addition to the t-tests we also calculated Pearson's correlation (shown in Table 5) between the features and the average system coverage score for each set. In the correlation analysis the input sets are not classified into easy or difficult but rather the real valued coverage scores are used directly. Overall, the features that were identified by the t-test as most descriptive of the differences between easy and difficult inputs were also the ones with higher correlations with real-valued coverage scores.

Our expectations in defining the features are confirmed by the correlation results. For example, systems have low coverage scores for sets with high-entropy vocabularies as indicated by the negative and high by absolute value correlation (-0.4256). Sets with high entropy are those in which there is little repetition within and across different articles, and for which it is subsequently difficult to determine what is the most important content. On the other hand, sets characterized by bigger KL divergence are easier—there the distribution of words is skewed compared to a general collection of articles, with important topic words occurring more often.

Easy to summarize sets are characterized by low entropy, small vocabulary, high average cosine and average topic signature overlaps, high KL divergence and a high percentage of the vocabulary consists of topic signature terms.

5 Classification results

We used the 192 sets from multi-document summarization DUC evaluations in 2002 (55 generic sets), 2003 (30 generic summary sets and 7 viewpoint sets) and 2004 (50 generic and 50 biography sets) to train and test a logistic regression classifier. The sets from all years were pooled together and evenly divided into easy and difficult inputs based on the average system coverage score for each set.

Table 6 shows the results from 10-fold cross validation. SIG is a classifier based on the six features identified as significant in distinguishing easy from difficult inputs based on a t-test comparison (Table 4). SIG+yt has two additional features: the year and the type of summarization input (generic, viewpoint and biographical). ALL is a classifier based on all 14 features defined in the previous section, and

feature	correlation
set entropy	-0.4256
KL divergence	0.3663
vocabulary size	-0.3610
% of sig. terms in vocab	0.3277
average sig. term overlap	0.2860
number of sentences	-0.2511
max topic signature overlap	0.2416
average cosine overlap	0.2244
number of signature terms	-0.1880
max cosine overlap	0.1337
min topic signature overlap	0.0401
min cosine overlap	0.0308
type-token ratio	-0.0276
% of words used only once	-0.0025

Table 5: Correlation between coverage score and feature values for the 29 DUC'01 100-word summaries.

features	accuracy	P	R	F
SIG	56.25%	0.553	0.600	0.576
SIG+yt	69.27%	0.696	0.674	0.684
ALL	61.45%	0.615	0.589	0.600
ALL+yt	65.10%	0.643	0.663	0.653

Table 6: Logistic regression classification results (accuracy, precision, recall and f-measure) for balanced data of 100-word summaries from DUC'02 through DUC'04.

ALL+yt also includes the year and task features.

Classification accuracy is considerably higher than the 50% random baseline. Using all features yields better accuracy (61%) than using solely the 6 significant features (accuracy of 56%). In both cases, adding the year and task leads to extra 3% net improvement. The best overall results are for the SIG+yt classifier with net improvement over the baseline equal to 20%. At the same time, it should be taken into consideration that the amount of training data for our experiments is small: a total of 192 sets. Despite this, the measures of input cohesiveness capture enough information to result in a classifier with above-baseline performance.

6 Conclusions

We have addressed the question of what makes the writing of a summary for a multi-document input difficult. Summary length is a significant factor, with all summarizers (people, machines and baselines) performing better at longer summary lengths. An exploratory analysis of DUC 2001 indicated that systems produce better summaries for cohesive inputs dealing with a clear topic (single event, subject and biographical sets) while non-cohesive sets about multiple events and opposing opinions are consistently of lower quality. We defined a number of features aimed at capturing input cohesiveness, ranging from simple features such as input length and size to more sophisticated measures such as input set entropy, KL divergence from a background corpus and topic signature terms based on log-likelihood ratio.

Generally, easy to summarize sets are characterized by low entropy, small vocabulary, high average cosine and average topic signature overlaps, high KL divergence and a high percentage of the vocabulary consists of topic signature terms. Experiments with a logistic regression classifier based on the features further confirms that input cohesiveness is predictive of the difficulty it will pose to automatic summarizers.

Several important notes can be made. First, it is important to develop strategies that can better handle non-cohesive inputs, reducing fluctuations in system performance. Most current systems are developed with the expectation they can handle any input but this is evidently not the case and more attention should be paid to the issue. Second, the interpretations of year to year evaluations can be affected. As demonstrated, the properties of the input have a considerable influence on summarization quality. If special care is not taken to ensure that the difficulty of inputs in different evaluations is kept more or less the same, results from the evaluations are not comparable and we cannot make general claims about progress and system improvements between evaluations. Finally, the presented results are clearly just a beginning in understanding of summarization difficulty. A more complete characterization of summarization input will be necessary in the future.

References

Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*.

David Carmel, Elad Yom-Tov, Adam Darlow, and Dan

Pelleg. 2006. What makes a query difficult? In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 390–397.

John Conroy, Judith Schlesinger, and Dianne O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of ACL, companion volume*.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR 2002)*, pages 299–306.

Ted Dunning. 1994. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Surabhi Gupta, Ani Nenkova, and Dan Jurafsky. 2007. Measuring importance and query relevance in topic-focused multi-document summarization. In *ACL'07, companion volume*.

BalaKrishna Kolluru and Yoshihiko Gotoh. 2005. On the subjectivity of human authored short summaries. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Finley Lacatusu, Andrew Hickl, Sanda Harabagiu, and Luke Nezda. 2004. Lite_gistexter at duc2004. In *Proceedings of the 4th Document Understanding Conference (DUC'04)*.

F. Lacatusu, A. Hickl, K. Roberts, Y. Shi, J. Bensley, B. Rink, P. Wang, and L. Taylor. 2006. Lcc's gistexter at duc 2006: Multi-strategy multi-document summarization. In *DUC'06*.

Chin-Yew Lin and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501.

Chin-Yew Lin and Eduard Hovy. 2003a. Automatic evaluation of summaries using n-gram co-occurance statistics. In *Proceedings of HLT-NAACL 2003*.

Chin-Yew Lin and Eduard Hovy. 2003b. The potential and limitations of automatic sentence extraction for summarization. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop*, pages 73–80.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *ACL Text Summarization Workshop*.

H. P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, B. Schiffman, and S. Teufel. 2001. Columbia multi-document summarization: Approach and evaluation. In *DUC'01*.

Kathleen McKeown, Regina Barzilay, David Evans, Vasleios Hatzivassiloglou, Judith Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the 2nd Human Language Technologies Conference HLT-02*.

Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *Proceedings of SIGIR*.

Dragomir Radev and Daniel Tam. 2003. Single-document and multi-document summary evaluation via relative utility. In *Poster session, International Conference on Information and Knowledge Management (CIKM'03)*.

Dragomir Radev, Hongyan Jing, Malgorzata Sty, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938.

Elad Yom-Tov, Shai Fine, David Carmel, and Adam Darlow. 2005. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 512–519.