A Compositional Context Sensitive Multi-document Summarizer: Exploring the Factors That Influence Summarization

Ani Nenkova
Stanford University
anenkova@stanford.edu

Lucy Vanderwende
Microsoft Research
lucyv@microsoft.com

Kathleen McKeown
Columbia University
kathy@cs.columbia.edu

ABSTRACT

The usual approach for automatic summarization is sentence extraction, where key sentences from the input documents are selected based on a suite of features. While word frequency often is used as a feature in summarization, its impact on system performance has not been isolated. In this paper, we study the contribution to summarization of three factors related to frequency: content word frequency, composition functions for estimating sentence importance from word frequency, and adjustment of frequency weights based on context. We carry out our analysis using datasets from the Document Understanding Conferences, studying not only the impact of these features on automatic summarizers, but also their role in human summarization. Our research shows that a frequency based summarizer can achieve performance comparable to that of state-of-the-art systems, but only with a good composition function; context sensitivity improves performance and significantly reduces repetition.

Categories and Subject Descriptors

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

General Terms

Measurement, Experimentation, Human Factors

Keywords

multi-document summarization, frequency, compositionality, context-sensitivity

1. INTRODUCTION

Most current automatic summarization systems rely on sentence extraction[1], where key sentences in the input documents are selected to form the summary. Even systems that

---

[1]A description of some most recent systems can be found

go beyond sentence extraction, reformulating or simplifying the text of the original articles, must decide which sentences should be simplified, compressed, fused together or rewritten [10, 11, 28, 2, 6]. Common approaches for identifying important sentences to include in the summary include training a binary classifier (e.g., [12]), training a Markov model (e.g., [4]), or directly assigning weights to sentences based on a variety of features and heuristically determined feature weights (e.g., [26, 14]). But the question of which components and features of automatic summarizers contribute most to their performance has largely remained unanswered [18]. In this paper, we examine several design decisions and the impact they have on the performance of generic multi-document summarizers of news. More specifically, we study the following issues:

Content word frequency. Word frequency is one feature that has been used in many summarization systems and originated in the earliest summarization research [17]. In this approach, content words such as nouns, verbs and adjectives serve as surrogates for the atomic units of meaning in text. While frequency has been used as a feature in many summarization systems, no study has isolated its impact on system performance. Only recently have large testsets for evaluation become available as a result of the annual Document Understanding Conference (DUC) run by NIST, which enable analysis of performance, and by the time DUC began, most systems were using a combination of features and not frequency alone. In this paper, we study the contribution of content word frequency in the input to system performance, showing that content word frequency also plays a role in human summarization behavior.

Choice of composition function. The frequency, and thus the importance, of content words can easily be estimated from the input to a summarizer. But is this enough to build a summarization system? Normally, a summarizer produces readable text as a summary, not a list of keywords, and thus it must estimate the importance of larger text units, typically sentences. A composition function needs to be chosen that will estimate the importance of a sentence as a function of the importance of the content words that appear in the sentence. There are many possibilities for the choice of composition function, and in Section 3 we will discuss three of them, showing that the choice can have a significant impact on the performance of the summarizer, ranging from close to baseline performance to overall state-of-the-art performance.

---

in the online proceedings of the Document Understanding Conference `http://duc.nist.gov`

Context sensitivity. The notion of importance is not static: it depends on what has been already said in a summary. Context adjustment is especially important for multi-document summarization (MDS), where the input consists of many articles on the same topic. Several articles might contain sentences expressing the same information. It is possible that they all get high importance weights and the summary will contain repetitive information. Avoiding repetition in the summary is a goal in summarization systems, since the very purpose of the summary is to reduce redundancy. We propose a method for context sensitivity and pinpoint its contribution to multi-document summarization performance. In Section 4 we show how context sensitivity adjustment improves content selection and reduces repetition in the summary.

We now proceed to a detailed discussion of these three aspects in the following sections.

2. FREQUENCY IN HUMAN SUMMARIES

One of the issues studied since the inception of automatic summarization in the 60s is that of human agreement [24]: different people choose different content for their summaries [27, 23, 19]. More recently, others have studied the degree of overlap between input documents and human summaries [5, 1]. The natural question that arises if we combine the two types of studies is whether features in the input can allow us to predict what content humans would choose in a summary, and what content they would agree on. If such predictors are identified, they could be used as features for content selection by an automatic system. In this section, we focus on frequency, investigating the association between content that appears frequently in the input, and the likelihood that it will be selected by a human summarizer for inclusion in a summary. This question is especially important for the multi-document summarization task, where the input consists of several articles on the same topic and usually contains a considerable amount of repetition of the same facts across documents. We first discuss the link between frequency in the input at the word level and the appearance of words in human summaries (Section 2.1), and then look at frequency at a semantic level, using manually identified semantic content units (Section 2.2).

2.1 Content word frequency and importance

In order to study how frequency influences human summarization choices, we used the 30 test sets for the multi-document summarization task from the large-scale common data set evaluation conducted within the DUC 2003. For each set, the input for summarization was available, along with four human abstracts for the input and the summaries produced by automatic summarizers that participated in the conference that year. Each of the inputs contained around 10 documents and the summaries were 100 words long. The counts for frequency in the input were taken over the concatenation of the documents in the input set.

The following instructions had been given to the human summarizers: "To write this summary, assume you have been given a set of stories on a news topic and that your job is to summarize them for the general news sections of the Washington Post. Your audience is the educated adult American reader with varied interestes and background in current and recent events."

2.1.1 Words frequent in the input appear in human summaries

We first turn to the question *Are content words that are very frequent in the input likely to appear in at least one of the human summaries?* We exclude stop words from consideration in this study, and use only nouns, verbs and adjectives. Table 1 shows the percentage of the N most frequent content words from the input documents that also appear in the human models, for $N = 5, 8, 12$. In order to compare how many of these matches are achieved by a good automatic summarizer, we picked one of the top performing summarizers and computed how many of the N most frequent words from the input documents appeared in its automatic summaries, and the numbers are shown in the second row of table 1. For example, the table shows that, across the 30 sets, 95% of the five most frequent content words in the input were also used in at least one of the summaries, while the automatic summarizer used only 84% (first column of table 1).

A comparison of this nature is helpful because in the commonly used intrinsic evaluations for summarization (discussed in more detail in Section 5), automatic summaries are evaluated by measuring their overlap with multiple human summaries/models.

Used by	5 most freq	8 most freq	12 most freq
Human	94.66%	91.25%	85.25%
Machine	84.00%	77.87%	66.08%

Table 1: Percentage of the N most frequent words from the input documents that appear in the four human models and in a state-of-the-art automatic summarizer (average across 30 input sets).

Two observations can be made about the table:

1. The high frequency words from the input are very likely to appear in the human models: the more frequent a word is in the input, the more likely it is that it will appear in a human summary. This confirms that frequency is one of the factors that impact a human's decision to include specific content in a summary. Probably observing frequency in the input indirectly helps the writers to resolve other constraints such as personal interests and background knowledge.

2. For the automatic summarizer, the trend to include more frequent words is preserved: the automatic summaries include 84% of the five most frequent words in the input, 78% of the 8 most frequent words, and 66% of the 12 most frequent. But the numbers are lower than those for the human summaries and the overlap between the machine summary and the human models can be improved if the inclusion of these most frequent words is targeted. As we will show later, it is possible to develop a summarizer that includes a percentage of most frequent words equivalent to that in *four* human summaries. Trying to maximize the number of matches with the human models is reasonable, since on average across the 30 sets, the machine summary contained 30 content words that did not match any word in a human model.[2]

---

[2]Even though no rigorous study of the issue has been done, it

Class C_i	Average $|C_i|$	Average frequency
C_4	7	31
C_3	11	14
C_2	24	9
C_1	82	5
C_0	1115	2

Table 2: C_i **(the first column) is the class of words that appear in** i **human summaries, Average** $|C_i|$ **(the second column) is the average size of class** C_i, **and the third column gives the average frequency of words in each class. The averages are computed for the 30 DUC'03 test sets.**

2.1.2 *Humans agree on words that are frequent in the input*

In the previous section we observed that the high frequency words in the input will tend to appear in *some* human model. But will high frequency words be words that the humans will agree on, and that will appear in *many* human summaries? In other words, we want to partition the words in the input into five classes C_n depending on how many human summaries they appear in, $n = 0...4$, and check if high class number is associated with higher frequency in the input for the words in the class. A word falls in C_0 if it does not appear in any of the human summaries, in C_1 if it appears in only one human summary and so on. Now we are interested to see how frequent the words in each class were in the respective input.

We found that, in fact, the words that human summarizers agreed to use in their summaries include the high frequency ones and the words that appear in only one human summary tend to be low frequency words as can be seen in table 2. The content words that were used by all four summarizers (in class C_4) had average frequency in the input equal to 31, while the words that never appeared in a human summary appeared on average about two times in the entire input of ten articles.

In the 30 sets of DUC 2003 data, the state-of-the-art machine summary contained 69% of the words appearing in all 4 human models and 46% of the words that appeared in 3 models. This indicates that high-frequency words, which human summarizers will tend to select and thus will be rewarded for example during automatic evaluation, are missing from the summary.

2.1.3 *Formalizing frequency: the multinomial model*

The findings from the previous sections suggest that frequency in the inputs is strongly indicative of whether a word will be used in a human summary. We start out with assessing the plausibility of a formal method capturing the relation between the occurrence of content words in the input and in summaries by modeling the appearance of words in the summary under a multinomial distribution estimated from the input. That is, for each word w in the input vocabulary, we associate a probability $p(w)$ for it to be emitted into a

can be considered that the content words that do not match any of the models describe "off-topic" events. This is consistent with the results from the quality evaluation of machine summaries in which human judges perceived more than half of the summary content to be "unnecessary, distracting or confusing."

Summarizer	Log-likelihood	Sum.	Log-like.
Human1:	-198.65	**S**ystem6:	-213.65
Human2:	-205.90	**H**uman9:	-215.65
Human3:	-205.91	**S**ystem7:	-215.92
Human4:	-206.21	**S**ystem8:	-216.04
Human5:	-206.37	**S**ystem9:	-216.20
System1:	-208.21	**S**ystem10:	-216.24
Human6:	-208.23	**S**ystem11:	-218.53
Human7:	-208.90	**S**ystem12:	-219.21
System2:	-210.14	**S**ystem13:	-220.31
System3:	-211.06	**S**ystem14:	-220.93
Human8:	-211.95	**S**ystem15:	-223.03
System4:	-212.57	**S**ystem16:	-225.20
System5:	-213.08	**H**uman10:	-227.17

Table 3: Average log-likelihood for the summaries of human and automatic summarizers in DUC'03. All summaries were truncated to 80 words to neutralize the effect of deviations from the required length of 100 words

summary. It is obvious that words with high frequency in the input will be assigned high emission probabilities.

The likelihood of a summary then is

$$L[sum; p(w_i)] = \frac{N!}{n_1!...n_r!} p(w_1)^{n_1} \cdot ... \cdot p(w_r)^{n_r} \qquad (1)$$

where N is the number of words in the summary, r is the number of unique words in the summary, $n_1 + ... + n_r = N$ and for each i, n_i is the number of times word w_i appears in the summary and $p(w_i)$ is the probability of w_i appearing in the summary estimated from the input documents. In order to confirm the hypothesis that human summaries have high likelihood under a multinomial model, we computed the log-likelihood $log[L(sum; p(w_i))]$ of all human and machine summaries from DUC'03 (see Table 3). There were 30 summaries from each system, and 12 summaries from each person. The log-likelihood is computed rather than the likelihood in order to avoid numeric problems such as underflow for very small probabilities. If human summaries have higher likelihood under the model than machine ones, we can conclude that a multinomial model captures more aspects of the human summarization process than of that of current automatic summarizers. And indeed: the log-likelihood of summaries produced by human summarizers were overall higher than for those produced by systems and the fact that the top five highest log-likelihood scores belong to humans indicate that some humans indeed employ a summarization strategy informed by frequency.[3]

2.2 Frequency of semantic content units

We established that high-frequency content words in the input are very likely to be used in human summaries, and that there will be a consensus about their inclusion in a summary between different human summarizers. But the co-occurrence of *words* in the inputs and the human summaries does not necessarily entail that the same *facts* have been covered. A better granularity for such investigation is the semantic content unit, an atomic fact expressed in a

[3]Other humans might have other strategies, such as giving maximum coverage of topics mentioned in the input, even those mentioned only once. Human10 appears to have such a strategy for example (after examination of his summaries).

text, such as the summary content units that form the basis of the pyramid method used for evaluation in the last DUC [19, 22]. In this annotation procedure, the content units are manually annotated[4], and expressions with the same meaning are linked together, even when there are differences in wording. For example, two documents can contain the sentences "Pinochet was arrested in the UK" and "Pinochet's arrest in Britain caused international controversy". While the wording is not exactly the same, both sentences express the content units *Pinochet was arrested* and *The arrest took place in Britain*.

Evans and McKeown [8] annotated 11 sets of DUC 2004 input documents and human written summaries for content units following the pyramid approach. Based on their annotation, we were able to measure how predictive the frequency of content units in the documents is for the selection of a content unit in a human summary. As in our study for words, we looked at the N most frequent content units in the inputs and calculated the percentage of these that appeared in any of the human summaries. Similarly to the case of words, of the 5 most frequent content units, 96% appeared in a human summary across the 11 sets. The respective percentages for the top 8 and top 12 content units were 92% and 85%. Thus content unit frequency is highly predictive for inclusion in a human summary, with the percentage of high frequency content units that are expressed in human summaries almost identical to the percentage for content words, presented in table 1.

Content units that are expressed in more human summaries, also occurred more often in the input, in agreement with the conclusion we drew from the analogous investigation on the word level.

In an additional experiment to confirm the hypothesis that frequency of content units is a predictive feature for summarization, we used the summarizer evaluation based on the 11 sets and reported in [7], and we computed the correlation between the weight of a content unit from the input documents (equal to the number of times the content unit was expressed in the input/its frequency) and the content unit weight from human summaries (equal to the number of summarizers that expressed the content unit in their summaries of the input). The Pearson's correlation coefficient between the input and human summaries weights is 0.64 (p-value=0), strongly indicating that content units that are repeated in several documents are likely to be picked in consensus by several humans and showing that frequency in the input helps predict human agreement in terms of content units. The lower than perfect correlation shows that there are other factors at play that influence human content selection decisions, which we do not find surprising at all and the discovery of which will be the focus of future work.

3. COMPOSITION FUNCTIONS

Now that we have shown that frequency is a good predictor of content in human summaries and that human summaries have higher likelihood under a multinomial model, how can we extend these empirical findings to building a summarizer? The question is not trivial: normally, only the frequency of content words can be easily obtained from the input, but how is the frequency of words to be combined in order to get an estimate for the importance of sentences, the

---

[4]Using a convenient visualization tool, DUCView.

usual units for extraction in summarization? We can define a family of summarizers, SUM_{CF}, where CF is the combination function yielding the importance of a sentence based on the words contained in that sentence. Different choices of CF will give different summarizers from the frequency based summarizer family. Below we outline the overall summarization algorithm and discuss possible choices of CF.

Context-sensitive frequency-based summarizer

Step 1 Compute the probability distribution over the words w_i appearing in the input, $p(w_i)$ for every i; $p(w_i) = \frac{n}{N}$, where n is the number of times the word appeared in the input, and N is the total number of content word tokens in the input. Only verbs, nouns, adjectives and numbers are considered in the computation of the probability distribution. Note that if part-of-speech tag were unavailable, we could use a simple stop word list in order to decide which words to count as content words.

Step 2 Assign an importance weight to each sentence S_j in the input as a function of the importance of its content words.

$Weight(S_j) = CF[p(w_i)]$ for $w_i \in S_j$

Step 3 Pick the best scoring sentence under the scoring function CF from the previous step.

Step 4 If the desired summary length has not been reached, go back to Step 2.

Different summarizers SUM_{CF} can be obtained by making different choices for the composition function CF. Three obvious candidates for CF are:

Product $(CF \equiv \prod)$ For this choice of CF
$Weight(S_j) = \prod_{w_i \in S_j} p(w_i)$

Average $(CF \equiv Avr)$ For this choice of CF
$Weight(S_j) = \frac{\sum_{w_i \in S_j} p(w_i)}{|\{w_i | w_i \in S_j\}|}$

Sum $(CF \equiv \sum)$ For this choice of CF
$Weight(S_j) = \sum_{w_i \in S_j} p(w_i)$

Each of these choices for CF leads to a different frequency based summarizer and we will see that the specific choice has a huge impact on the performance of the summarizer; not all frequency-based summarizers perform well.

How does a summarizer SUM_{CF} do in terms of inclusion of top frequency words compared to humans and other top performing systems? Table 4 shows the percentage of the N most frequent words from the DUC'03 documents that also appear in SUM_{Avr} summaries. As expected, these are much higher than the percentages for the non-frequency oriented machine summarizer; moreover, they are even higher than in all four human models taken together.

4. CONTEXT ADJUSTMENT

Using frequency alone to determine summary content in multi-document summarization will result in a repetitive summary. We can adjust the algorithm to account for information included so far by adding Step 3.5, shown below.

Used by	5 most freq	8 most freq	12 most freq
Human	94.66%	91.25%	85.25%
Machine	84.00%	77.87%	66.08%
$\mathbf{SUM}_{Avr}$	96.00%	95.00%	90.83%

Table 4: Percentage of the N most frequent words from the input documents that appear in one of the four human models, a state-of-the-art machine summarizer and SUM_{Avr}, a new machine summarizer based on frequency that uses the average as a composition function.

System	# of sentences	Sentences per summary
$\text{SUM}_{\prod}$	270	5.40
SUM_{Avr}	223	4.46
$\text{SUM}_{\sum}$	155	3.10

Table 5: Number of sentences in systems' summaries: the choice of composition function CF affects systems' preference to longer or shorter sentences and SUM_{Avr} is the more balanced one.

Step 3.5 For each word w_i in the sentence chosen at step 3, update its probability by setting it to a very small number close to 0. Here we used 0.0001 for this number.

It serves a threefold purpose:

1. It gives the summarizer sensitivity to context. The notion of what is most important to include in the summary changes depending on what information has already been included in the summary.

2. By updating the probabilities in this intuitive way, we also allow words with initially low probability to have higher impact on the choice of subsequent sentences. If we look back at table 2, we see that this is a reasonable goal, since the large class of words that were expressed only in one model were not that frequent; that is, content that humans will not necessarily agree on, but is still good for inclusion, is not characterized by high frequency.

3. The update of word probability gives a natural way to deal with the redundancy in the multi-document input. In fact, in terms of content units, the inclusion of the same unit twice in the same summary is rather improbable. As we see in the following evaluation section, no further checks for duplication seem to be necessary.

In the next section, we evaluate the algorithm both with and without step 3.5, showing that when it is removed from the algorithm, the summarizer does worse on content selection and there is a substantial increase in information repetition in the summary.

5. EVALUATION RESULTS

To evaluate the performance of the three SUM_{CF} summarizers, both with and without context sensitive adjustment, we use the test data from two large common data set evaluation initiatives—the 50 test sets for multi-document summarization task for DUC 2004 and the common test set provided in the 2005 Machine Translation and Summarization Evaluation (MSE) initiative. Both tasks were to produce a generic 100-word summary of several related articles, but in the MSE task some of the input consisted of machine translated text.

Document Understanding Conference

We used the data from the 2003 DUC conference for development and the data from the 2004 DUC as test data,

which we report on here. We tested the SUM_{CF} family of summarizers on the 50 sets from the generic summary task in 2004 DUC.

Even before analysis of quantitative metrics, we can see that the choice of combination function CF has a significant impact on summarizer performance. One would expect that the probabilistic summarizer $\text{SUM}_{\prod}$ would favor shorter sentences because as the sentence gets longer, their overall probability involves the multiplication of more word probabilities (numbers between 0 and 1) and thus overall longer sentences will have lower probability. Exactly the opposite would be expected from $\text{SUM}_{\sum}$, which assigns sentences a weight equal to the sum of probabilities of the words in the sentence. The more words there are in the sentence, the higher the sentence weight will tend to be. SUM_{Avr} is a compromise between the two extremes. To confirm this intuition about the behavior of the summarizers depending on the choice of CF, we looked at the length in sentences of the summaries that they produced. Table 5 shows the number of sentences across the 50 summaries produced by each of the systems. Our intuition is confirmed, with $\text{SUM}_{\sum}$ producing summaries of about three sentences and $\text{SUM}_{\prod}$ getting about five sentences per summary, for the same size in words. The average human summary for the same topics has around four sentences, close that for SUM_{Avr}.

For the evaluation, we use the ROUGE-1 automatic metric, which has been shown to correlate well with human judgments based on comparison with a single model [15, 13] and which was found to have one of the best correlations with human judgment on the DUC 2004 data [21] among the several possible automatic metrics. In addition, we report the ROUGE-2 and ROUGE-SU4 metrics, which were used as official automatic evaluation metrics for MSE 2005 and DUC 2005.

The results are obtained with ROUGE version 1.5.5 with the settings used for DUC 2005 (with *-s* option for removing stopwords for ROUGE-1).[5]

All summaries were truncated to 100 words (space delimited tokens) for the evaluation, as is normally done in DUC evaluations. The first column of table 6 also lists the number of words in the 50 summaries in the test set. Some systems did not generate the longest possible summary. Peer 120 was an extreme example, producing summaries with average length of 78 words. But the impact of peer summary length on the final ranking of the systems is unlikely to be big, since most systems produced summaries very close to the required 100 word limit.

An approximate result on determining which differences in scores are significant can be obtained by comparing the 95% confidence intervals for each mean. Significant differences

---

[5]The exact parameters we used were *-n 2 -x -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -d*

SYSTEM	ROUGE-1	ROUGE-2	ROUGE-SU4
peer 65 (4988)	0.305 (0.289; 0.320)	0.089 (0.081; 0.098)	0.130 (0.123; 0.137)
peer 34 (4954)	0.287 (0.271; 0.305)	0.074 (0.065; 0.083)	0.121 (0.113; 0.129)
peer 102 (4951)	0.285 (0.268; 0.303)	0.084 (0.076; 0.091)	0.126 (0.119; 0.132)
SUM$_\sum$ (5000)	**0.283 (0.267; 0.300)**	**0.079 (0.072; 0.087)**	**0.122 (0.115; 0.129)**
peer 124 (4988)	0.282 (0.265; 0.300)	0.081 (0.073; 0.088)	0.123 (0.116; 0.131)
SUM$_{Avr}$ (5000)	**0.280 (0.265; 0.297)**	**0.076 (0.069;0.084)**	**0.121 (0.115; 0.127)**
peer 44 (4854)	0.273 (0.256; 0.290)	0.076 (0.067; 0.084)	0.119 (0.111; 0.126)
peer 81 (4994)	0.268 (0.251; 0.285)	0.078 (0.070; 0.087)	0.121 (0.113; 0.128)
peer 55 (4971)	0.262 (0.247; 0.280)	0.069 (0.062; 0.077)	0.114 (0.107; 0.121)
peer 93 (4612)	0.253 (0.235; 0.271)	0.072 (0.066; 0.080)	0.107 (0.101; 0.114)
SUM$_{AvrNoAdjust}$ (5000)	**0.252 (0.235; 0.269)**	**0.075 (0.069; 0.083)**	**0.116 (0.108; 0.124)**
peer 120 (3903)	0.251 (0.231; 0.271)	0.077 (0.068; 0.085)	0.108 (0.099; 0.117)
peer 117 (4997)	0.238 (0.221; 0.257)	0.057 (0.051; 0.063)	0.107 (0.100; 0.113)
peer 140 (5000)	0.239 (0.219; 0.260)	0.068 (0.060; 0.076)	0.108 (0.101; 0.116)
peer 11 (4172)	0.239 (0.218; 0.259)	0.071 (0.062; 0.080)	0.105 (0.096; 0.114)
peer 138 (5000)	0.230 (0.211; 0.253)	0.069 (0.061; 0.077)	0.106 (0.098; 0.113)
SUM$_\prod$ (5000)	**0.227 (0.210; 0.245)**	**0.058 (0.050; 0.065)**	**0.104 (0.097; 0.110)**
Baseline (4899)	*0.202 (0.183; 0.221)*	*0.061 (0.052; 0.070)*	*0.098 (0.092; 0.106)*
peer27 (4686)	0.185 (0.166; 0.204)	0.046 (0.039; 0.055)	0.090 (0.083; 0.098)
peer123 (4338)	0.189 (0.173; 0.206)	0.049 (0.043; 0.056)	0.090 (0.084; 0.096)
peer 111 (5000)	0.063 (0.053; 0.073)	0.016 (0.013; 0.019)	0.057 (0.053; 0.061)

Table 6: DUC'04 ROUGE-1, ROUGE-2 and ROUGE-SU4 stemmed, stop-words removed for ROUGE-1 test set scores and their 95% confidence intervals for participating systems, the baseline, and SUM$_{CF}$.

are those where the confidence intervals for the estimates of the means for the two systems either do not overlap at all, or where the two intervals overlap but neither contains the best estimate for the mean of the other, though [25] warns that the latter approach may indicate significance more often than it should.

Table 6 also shows scores for the 16 other participating systems from DUC 2004, and the baseline, which was selecting the beginning of the latest article as a summary.

Several conclusions can be drawn from the table:

Comparison between SUM$_{CF}$ summarizers: All three SUM$_{CF}$ summarizers use word frequency in the input as a feature but have a different composition function CF to assign weights to sentences. SUM$_\prod$ is a probabilistic summarizer and the weight it assigns to each sentence is in fact the probability of the sentence. SUM$_{Avr}$ and SUM$_\sum$ assign to sentences weight equal to the average and the sum of the probabilities of the words in the sentence respectively. For these two latter summarizers, the raw frequency of words could be used instead of word probabilities. For all three automatic metrics, SUM$_\prod$ is *significantly worse* than SUM$_\sum$ and SUM$_{Avr}$ and is in fact very close to baseline performance. SUM$_{Avr}$ and SUM$_\sum$ are almost identical in terms of ROUGE scores.

The effect of context adjustment: In the table we have listed the automatic scores for SUM$_{AvrNoAdjust}$. This is the summarizer for which the composition function $CF \equiv Avr$, but without *Step 3.5* from the summarization algorithm, which is responsible for adjusting the weights for words that appear in sentences already chosen for inclusion in the summary. All three metrics indicate that the content selection capability of the summarizer is affected by the removal of the context adjustment step. According to ROUGE-1, removing the context adjustment leads to *significantly* lower results, while for the other two metrics the deterioration is not significant. In order to assess how much *Step 3.5* affected the occurance of repetition in the summaries, we analyzed 10 of the produced summaries for repeated content units. There were 3 repeated content units

in the SUM$_{Avr}$ summaries, and 13 repeated content units in the SUM$_{AvrNoAdjust}$ summaries, which is a substantial increase.

Comparison with other DUC systems SUM$_\sum$ and SUM$_{Avr}$ perform extremely well compared to the other DUC 2004 systems. Peer 65 is the only system that significantly outperforms them, while ten (more than half) of the other systems are significantly worse. It is worth noting that peer 65 is a supervised HMM system [4], requring training data and parameter adjustment, while the SUM$_{CF}$ summarizers are non-supervised and totally data-driven. In sum, the SUM$_{CF}$ summarizers are about as good as the best DUC 2004 participants.

Overall, SUM$_{Avr}$ is the best of the SUM$_{CF}$ family in balancing content selection scores and sentence length preference, and this is the summarizer we choose for later comparisons. Its sentence selection scores are comparable to that of the best DUC 2004 summarizers, it has most success in avoiding repetition in the summary from the frequency summarizer family, and it is least sensitive to the influence of sentence length on the sentence weight.

Machine translation and summarization evaluation 2005

In April 2005, a multi-document summarization evaluation task was conducted as part of the Machine Translation and Summarization Workshop at ACL.[6] The task was to produce a 100-word summary from multi-document inputs consisting of a mixture of English documents and machine translations to English of Arabic documents on the same topic. Some summarizers were modified for this task to use redundancy to correct errors in the machine translations, or to avoid MT text altogether and choose only sentences from the English input.

We ran SUM$_{Avr}$ without any modifications to account for the non-standard input [29]. The light-weight version of the summarizer was run, which did not require part of speech

---

[6]`http://www.isi.edu/~cyl/MTSE2005/MLSummEval.html`

system	pyramid	R-2	R-SU4	repetition
1	0.52859	0.13076	0.15670	1.4
28	0.48926	0.16036***	0.18627***	3.4***
19	0.45852	0.11849	0.14971***	1.3
SUM_{Avr}	*0.45274*	*0.12678*	*0.15938*	*0.6*
10	0.44254	0.13038	0.16568	1.2
16	0.45059	0.13355	0.16177	0.9
13	0.43429	0.08580***	0.11141***	0.4
25	0.39823	0.11678	0.15079	2.7***
4	0.37297	0.12010	0.15394	4.1***
7	0.37159	0.09654***	0.13593	0.4

Table 7: Results from the MSE evaluation. Pyramid scores and duplication is computed for 10 test sets, automatic scores for all 25 test sets. Numbers flagged by "*" are significantly different from the results form SUM_{Avr}. For repetition, higher numbers are worse, indicating that there was more repetition in the summary.**

tags and which excluded stop words from a given stop word list.

The official evaluation metrics adopted for the workshop were the manual pyramid score, ROUGE-2 (the bigram overlap metric) and ROUGE-SU4 (skip bigram). The skip bigram metric measures the occurrence of a pair of words in their original sentence order, permitting up to four intervening words. The metric was originally proposed for machine translation evaluation and was shown to correlate well with human judgments both for machine translation and for summarization [13, 16].

The pyramid method was used to evaluate only 10 of the test sets, while the automatic metrics were applied to all 25 test sets. The average results for each peer for the three metrics is shown in table 7. For the manual pyramid scores, none of the differences between systems were significant according to a paired t-test at the 5% level of significance. This is not surprising, given the small number of test points. There were only three peers with average scores larger than that of SUM_{Avr}, and six systems with lower average pyramid performance. We again see that SUM_{Avr} is competitive in comparison with other, more sophisticated, MDS systems in terms of content selection and is one of the best systems in avoiding repetion in the summaries.

For the automatic metrics, significance was based again on the 95% confidence interval provided by ROUGE. One system was significantly better than SUM_{Avr}, and for each of the automatic metrics there were two systems that were significantly worse than SUM_{Avr}. The rest of the differences were not significant. In table 7, results that are significantly different from those for SUM_{Avr} are flagged by "***".

During the annotation for the pyramid scoring, the content units that were repeated in an automatic summary were marked up: we include in the results table the average number of repeated SCUs per summary for all systems. SUM_{Avr} was one of the systems with the lowest amount of repetition in its summaries, with three of the other peers including significantly more repetitive information. These results confirm our intuition that the weight update of words to adjust for context is sufficient for dealing with duplication removal problems. This experiment also confirms that SUM_{Avr} is a robust summarizer with good performance.

6. RELATED WORK

Maximal Marginal Relevance (MMR) is the method for redundancy removal mentioned most often in the context of summarization research. The method was first introduced in [3] and was applied for multi-document summarization in [9]. The MMR approach was developed primarily for information retrieval and query-focused summarization, and gives a summarizer sensitivity to context by reweighting sentences using a linear combination of the similarity between the sentence and *1)* the query and *2)* the summary sentences already selected in the summary. The best sentence is considered the one that is most similar to the query and least similar to the text that is already in the summary. In [9], the technique was used to create multi-document extracts of 25 sets of 10 articles each. The evaluation was done by computing the cosine similarity between the extract and a human model extract for the same set. In this setting, extracts produced using MMR and those not using the technique received the same evaluation score, and thus the usefulness of the technique could not be demonstarted. Many systems use the MMR idea for generic multi-document summarization,[7] where no user query is available, by setting a single paramater for similarity and rejecting all sentences that have similarity with the already chosen part of the summary that exceeds this predefined treshold. An evaluation of how changing this paramater influences the quality of the summaries has not been reported. In addition to this similarity parameter, the similarity measure that is used makes a difference for the success in duplication removal, as reported in [20], who focused on the study of different similarity metrics for duplication removal.

7. CONCLUSIONS

Our analysis using the DUC datasets shows that frequency has a powerful impact on the performance of summarization systems, provided that a good composition function is used. Our results show that averaging word probabilities yields a system that performs comparably to other state-of-the-art systems and that outperfoms many of the participating systems. When context is taken into account and probabilities are adjusted when the word has already appeared in the summary, performance based on content shows an improvement, but more importantly, repetition in the summary significantly decreases.

These results suggest that the more complex combination of features used by state-of-the-art systems today may not be necessary and the contribution of such features needs to be precisely isolated. They highlight the fact that composition plays an important role in performance, but is an unknown for most state-of-the-art systems, who often do not report the composition function that was used. Furthermore, they demonstrate that repetition can be reduced within the same frequency-based model.

It is worth noting that the presented summarization algorithm uses frequency in a greedy way, choosing the current best sentence at each iteration. Such an approach does not take advantage of the result we demostrated that human summaries tend to have high likelihood under a multinomial model. This fact could be used in a global optimization algorithm, possibly leading to better results.

---

[7]See for example the online DUC 2004 proceeding

8. REFERENCES

[1] M. Banko and L. Vanderwende. Using n-grams to understand the nature of summaries. In *Proceedings of HLT/NAACL'04*, 2004.

[2] R. Barzilay and K. McKeown. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3), 2005.

[3] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 335–336, 1998.

[4] J. Conroy, J. Schlesinger, J. Goldstein, and D. O'Leary. Left-brain/right-brain multi-document summarization. In *Proceedings of the 4th Document Undersatnding Conference (DUC'04)*, 2004.

[5] T. Copeck and S. Szpakowicz. Vocabulary agreement among model summaries and source documents. In *Proceedings of the Document Understanding Conference DUC'04*, 2004.

[6] H. Daumé III and D. Marcu. Bayesian multi-document summarization at mse. In *Proceedings of the Workshop on Multilingual Summarization Evaluation (MSE)*, Ann Arbor, MI, June 29 2005.

[7] D. K. Elson. Project logline: Rhetorical categorization for multidocument news summarization. Master's thesis, Columbia University, 2005.

[8] D. K. Evans and K. McKeown. Identifying similarities and differences across english and arabic news. In *Proceedings of the International Conference on Intelligence Analysis*, 2005.

[9] J. Goldstein, V. Mittal, J. Carbonell, and J. Callan. Creating and evaluating multi-document sentence extract summaries. In *CIKM '00: Proceedings of the ninth international conference on Information and knowledge management*, pages 165–172, 2000.

[10] H. Jing and K. McKeown. Cut and paste based text summarization. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'00)*, 2000.

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

[12] J. Kupiec, J. Perersen, and F. Chen. A trainable document summarizer. In *Research and Development in Information Retrieval*, pages 68–73, 1995.

[13] C.-Y. Lin. Rouge: a package for automatic evaluation of summaries. In *Proceedings of the Workshop in Text Summarization, ACL'04*, 2004.

[14] C.-Y. Lin and E. Hovy. Automated multi-document summarization in neats. In *Proceedings of the Human Language Technology Conference (HLT2002)*, 2002.

[15] C.-Y. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurance statistics. In *Proceedings of HLT-NAACL 2003*, 2003.

[16] C.-Y. Lin and F. J. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, 2004.

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

[18] D. Marcu and L. Gerber. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*, 2001.

[19] A. Nenkova and R. Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT/NAACL 2004*, 2004.

[20] E. Newman, W. Doran, N. Stokes, J. Carthy, and J. Dunnion. Comparing redundancy removal techniques for multi-document summarisation. In *Proceedings of STAIRS*, pages 223–228, 2004.

[21] P. Over and J. Yen. An introduction to duc 2004 intrinsic evaluation of generic news text summarization systems. In *Proceedings of DUC 2004*, 2004.

[22] R. Passonneau, A. Nenkova, K. McKeown, and S. Sigleman. Pyramid evaluation ot duc 2005. In *Proceedings of the Document Understanding Conference (DUC'05)*, 2005.

[23] D. Radev, S. Teufel, H. Saggion, and W. Lam. Evaluation challenges in large-scale multi-document summarization. In *ACL*, 2003.

[24] G. J. Rath, A. Resnick, and R. Savage. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. *American Documentation*, 2(12):139–208, 1961.

[25] N. Schenker and J. Gentleman. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician*, 55(3):182–186, 2001.

[26] B. Schiffman, A. Nenkova, and K. McKeown. Experiments in multidocument summarization. In *Proceedings of the Human Language Technology Conference*, 2002.

[27] H. van Halteren and S. Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*, 2003.

[28] L. Vanderwende, M. Banko, and A. Menezes. Event-centric summary generation. In *Proceedings of the Document Understanding Conference (DUC'04)*, 2004.

[29] L. Vanderwende and H. Suzuki. Frequency-based summarizer and a language modeling extention. In *MSE 2005 common data task evaluation*, 2005.