Query-Focused Summaries or Query-Biased Summaries ?

Rahul Katragadda
Language Technologies Research Center
IIIT Hyderabad
rahul_k@research.iiit.ac.in

Vasudeva Varma
Language Technologies Research Center
IIIT Hyderabad
vv@iiit.ac.in

Abstract

In the context of the Document Understanding Conferences, the task of Query-Focused Multi-Document Summarization is intended to improve agreement in content among human-generated model summaries. *Query-focus* also aids the automated summarizers in directing the summary at specific topics, which may result in better agreement with these model summaries. However, while query focus correlates with performance, we show that high-performing automatic systems produce summaries with disproportionally higher query term density than human summarizers do. Experimental evidence suggests that automatic systems heavily rely on query term occurrence and repetition to achieve good performance.

1 Introduction

The problem of automatically summarizing text documents has received a lot of attention since the early work by Luhn (Luhn, 1958). Most of the current automatic summarization systems rely on a sentence extractive paradigm, where key sentences in the original text are selected to form the summary based on the clues (or heuristics), or learning based approaches.

Common approaches for identifying key sentences include: training a binary classifier (Kupiec et al., 1995), training a Markov model or CRF (Conroy et al., 2004; Shen et al., 2007) or directly assigning weights to sentences based on a variety of features and heuristically determined feature weights (Toutanova et al., 2007). But, the question of which components and features of automatic summarizers contribute most to their performance has largely remained unanswered (Marcu and Gerber, 2001), until Nenkova et al. (Nenkova et al., 2006) explored the contribution of frequency based measures. In this paper, we examine the role a *query* plays in automated multi-document summarization of newswire.

One of the issues studied since the inception of automatic summarization is that of human agreement: different people choose different content for their summaries (Rath et al., 1961; van Halteren and Teufel, 2003; Nenkova et al., 2007). Later, it was assumed (Dang, 2005) that having a question/query to provide focus would improve agreement between any two human-generated model summaries, as well as between a model summary and an automated summary. Starting in 2005 until 2007, a *query-focused multi-document summarization* task was conducted as part of the annual Document Understanding Conference. This task models a real-world complex question answering scenario, where systems need to synthesize from a set of 25 documents, a brief (250 words), well organized fluent answer to an information need.

Query-focused summarization is a topic of ongoing importance within the summarization and question answering communities. Most of the work in this area has been conducted under the guise of "*query-focused multi-document summarization*", "*descriptive question answering*", or even "*complex question answering*".

In this paper, based on structured empirical evaluations, we show that most of the systems participating in DUC's Query-Focused Multi-Document Summarization (QF-MDS) task have been query-biased in building extractive summaries. Throughout our discussion, the term *'query-bias'*, with respect to a sentence, is precisely defined to mean that the sentence has at least one query term within it. The term *'query-focus'* is less precisely defined, but is related to the cognitive task of focusing a summary on the query, which we assume humans do naturally. In other words, the human generated model summaries are assumed to be query-focused.

Here we first discuss *query-biased* content in Summary Content Units (SCUs) in Section 2 and then in Section 3 by building formal models on *query-bias* we discuss why/how automated systems are query-biased rather than being query-focused.

2 Query-biased content in Summary Content Units (SCUs)

Summary content units, referred as SCUs hereafter, are semantically motivated subsentential units that are variable in length but not bigger than a sentential clause. SCUs are constructed from annotation of a collection of human summaries on a given document collection. They are identified by noting information that is repeated across summaries. The repetition is as small as a modifier of a noun phrase or as large as a clause. The evaluation method that is based on overlapping SCUs in human and automatic summaries is called the

Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pages 105–108,
Suntec, Singapore, 4 August 2009.

```
<document name="APW20000824.0204">
<line>A lawyer who specializes in bankrupting hate groups is going after
 the Aryan Nations, whose compound in the Idaho woods has served as a
 clubhouse for some of America's most violent racists.</line>
<line>In a lawsuit that goes to trial Monday, attorney Morris Dees of the
 Southern Poverty Law Center is representing a mother and son who were
attacked by security guards for the white supremacist group.
<annotation scu-count="1" sum-count="8" sums="13,14,15,23,24,29,30,9">
<scu uid="24" label="SPLC takes legal action against civil rights abuses"
 weight="3"/></annotation></line>
<line>The victims are suing the Aryan Nations and founder Richard Butler.
<annotation scu-count="0" sum-count="1" sums="29"/></line>
```

Figure 1: SCU annotation of a source document.

pyramid method (Nenkova et al., 2007).

The University of Ottawa has organized the pyramid annotation data such that for some of the sentences in the original document collection, a list of corresponding content units is known (Copeck et al., 2006). A sample of an SCU mapping from topic *D0701A* of the DUC 2007 QF-MDS corpus is shown in Figure 1. Three sentences are seen in the figure among which two have been annotated with system IDs and SCU weights wherever applicable. The first sentence has not been picked by any of the summarizers participating in Pyramid Evaluations, hence it is unknown if the sentence would have contributed to any SCU. The second sentence was picked by 8 summarizers and that sentence contributed to an SCU of weight 3. The third sentence in the example was picked by one summarizer, however, it did not contribute to any SCU. This example shows all the three types of sentences available in the corpus: unknown samples, positive samples and negative samples.

We extracted the positive and negative samples in the source documents from these annotations; types of second and third sentences shown in Figure 1. A total of 14.8% sentences were annotated to be either positive or negative. When we analyzed the positive set, we found that 84.63% sentences in this set were *query-biased*. Also, on the negative sample set, we found that 69.12% sentences were *query-biased*. That is, on an average, 76.67% of the sentences picked by any automated summarizer are *query-biased*. On the other hand, for human summaries only 58% sentences were *query-biased*. All the above numbers are based on the DUC 2007 dataset shown in **boldface** in Table 1 [1].

There is one caveat: The annotated sentences come only from the summaries of systems that participated in the pyramid evaluations. Since only 13 among a total 32 participating systems were evaluated using pyramid evaluations, the dataset is limited. However, despite this small issue, it is very clear that at least those systems that participated in pyramid evaluations have been biased towards query-terms, or at least, they have been better at correctly identifying important sentences from the query-biased sentences than from query-unbiased sentences.

[1]We used DUC 2007 dataset for all experiments reported.

3 Formalizing *query-bias*

Our search for a formal method to capture the relation between occurrence of query-biased sentences in the input and in summaries resulted in building binomial and multinomial model distributions. The distributions estimated were then used to obtain the likelihood of a query-biased sentence being emitted into a summary by each system.

For the DUC 2007 data, there were 45 summaries for each of the 32 systems (labeled 1-32) among which 2 were baselines (labeled 1 and 2), and 18 summaries from each of 10 human summarizers (labeled A-J). We computed the log-likelihood, $\log(L[summary;p(C_i)])$, of all human and machine summaries from DUC'07 query focused multi-document summarization task, based on both distributions described below (see Sections 3.1, 3.2).

3.1 The binomial model

We represent the set of sentences as a binomial distribution over type of sentences. Let C_0 and C_1 denote the sets of sentences without and with query-bias respectively. Let p(C_i) be the probability of emitting a sentence from a specified set. It is also obvious that query-biased sentences will be assigned lower emission probabilities, because the occurrence of query-biased sentences in the input is less likely. On average each topic has 549 sentences, among which 196 contain a query term; which means only 35.6% sentences in the input were query-biased. Hence, the likelihood function here denotes the likelihood of a summary to contain non query-biased sentences. Humans' and systems' summaries must now constitute low likelihood to show that they rely on *query-bias*.

The likelihood of a summary then is :

$$L[summary; p(C_i)] = \frac{N!}{n_0!n_1!} p(C_0)^{n_0} p(C_1)^{n_1} \quad (1)$$

Where N is the number of sentences in the summary, and n_0 + n_1 = N; n_0 and n_1 are the cardinalities of C_0 and C_1 in the summary. Table 2 shows various systems with their ranks based on ROUGE-2 and the average log-likelihood scores. The ROUGE (Lin, 2004) suite of metrics are n-gram overlap based metrics that have been shown to highly correlate with human evaluations on content responsiveness. ROUGE-2 and ROUGE-SU4 are the official ROUGE metrics for evaluating *query-focused multi-document summarization* task since DUC 2005.

3.2 The multinomial model

In the previous section (Section 3.1), we described the binomial model where we classified each sentence as being *query-biased* or not. However, if we were to quantify the amount of *query-bias* in a sentence, we associate each sentence to one among k possible classes leading to a multinomial distribution. Let $C_i \in$

Dataset	total	positive	biased positive	negative	biased negative	% bias in positive	% bias in negative
DUC 2005	24831	1480	1127	1912	1063	76.15	55.60
DUC 2006	14747	1047	902	1407	908	86.15	71.64
DUC 2007	***12832***	***924***	***782***	***975***	***674***	***84.63***	***69.12***

Table 1: Statistical information on counts of *query-biased* sentences.

ID	rank	LL	ROUGE-2	ID	rank	LL	ROUGE-2	ID	rank	LL	ROUGE-2
1	31	-1.9842	0.06039	***J***		***-3.9465***	0.13904	24	4	-5.8451	0.11793
C		-2.1387	0.15055	***E***		***-3.9485***	0.13850	9	12	-5.9049	0.10370
16	32	-2.2906	0.03813	10	28	-4.0723	0.07908	14	14	-5.9860	0.10277
27	30	-2.4012	0.06238	21	22	-4.2460	0.08989	5	23	-6.0464	0.08784
6	29	-2.5536	0.07135	G		-4.3143	0.13390	4	3	-6.2347	0.11887
12	25	-2.9415	0.08505	25	27	-4.4542	0.08039	20	6	-6.3923	0.10879
I		-3.0196	0.13621	B		-4.4655	0.13992	29	2	-6.4076	0.12028
11	24	-3.0495	0.08678	19	26	-4.6785	0.08453	3	9	-7.1720	0.10660
28	16	-3.1932	0.09858	26	21	-4.7658	0.08989	8	11	-7.4125	0.10408
2	18	-3.2058	0.09382	23	7	-5.3418	0.10810	17	15	-7.4458	0.10212
D		***-3.2357***	0.17528	30	10	-5.4039	0.10614	13	5	-7.7504	0.11172
H		***-3.4494***	0.13001	7	8	-5.6291	0.10795	32	17	-8.0117	0.09750
A		***-3.6481***	0.13254	18	19	-5.6397	0.09170	22	13	-8.9843	0.10329
F		***-3.8316***	0.13395	15	1	-5.7938	0.12448	31	20	-9.0806	0.09126

Table 2: Rank, Averaged log-likelihood score based on **binomial model**, true ROUGE-2 score for the summaries of various systems in DUC'07 *query-focused multi-document summarization* task.

ID	rank	LL	ROUGE-2	ID	rank	LL	ROUGE-2	ID	rank	LL	ROUGE-2
1	31	-4.6770	0.06039	10	28	-8.5004	0.07908	5	23	-14.3259	0.08784
16	32	-4.7390	0.03813	G		-9.5593	0.13390	9	12	-14.4732	0.10370
6	29	-5.4809	0.07135	***E***		***-9.6831***	0.13850	22	13	-14.8557	0.10329
27	30	-5.5110	0.06238	26	21	-9.7163	0.08989	4	3	-14.9307	0.11887
I		-6.7662	0.13621	***J***		***-9.8386***	0.13904	18	19	-15.0114	0.09170
12	25	-6.8631	0.08505	19	26	-10.3226	0.08453	14	14	-15.4863	0.10277
2	18	-6.9363	0.09382	B		-10.4152	0.13992	20	6	-15.8697	0.10879
C		-7.2497	0.15055	25	27	-10.7693	0.08039	32	17	-15.9318	0.09750
H		***-7.6657***	0.13001	29	2	-12.7595	0.12028	7	8	-15.9927	0.10795
11	24	-7.8048	0.08678	21	22	-13.1686	0.08989	17	15	-17.3737	0.10212
A		***-7.8690***	0.13254	24	4	-13.2842	0.11793	8	11	-17.4454	0.10408
D		***-8.0266***	0.17528	30	10	-13.3632	0.10614	31	20	-17.5615	0.09126
28	16	-8.0307	0.09858	23	7	-13.7781	0.10810	3	9	-19.0495	0.10660
F		***-8.2633***	0.13395	15	1	-14.2832	0.12448	13	5	-19.3089	0.11172

Table 3: Rank, Averaged log-likelihood score based on **multinomial model**, true ROUGE-2 score for the summaries of various systems in DUC'07 *query-focused multi-document summarization* task.

$\{C_0, C_1, C_2, \ldots, C_k\}$ denote the k levels of *query-bias*. C_i is the set of sentences, each having i query terms.

The number of sentences participating in each class varies highly, with C_0 bagging a high percentage of sentences (64.4%) and the rest $\{C_1, C_2, \ldots, C_k\}$ distributing among themselves the rest 35.6% sentences. Since the distribution is highly-skewed, distinguishing systems based on log-likelihood scores using this model is easier and perhaps more accurate. Like before, Humans' and systems' summaries must now constitute low likelihood to show that they rely on *query-bias*.

The likelihood of a summary then is :

$$L[summary; p(C_i)] = \frac{N!}{n_0! n_1! \cdots n_k!} p(C_0)^{n_0} p(C_1)^{n_1} \cdots p(C_k)^{n_k} \tag{2}$$

Where N is the number of sentences in the summary, and $n_0 + n_1 + \cdots + n_k$ = N; n_0, n_1,$\cdots$,n_k are respectively the cardinalities of C_0, C_1, $\cdots$,C_k, in the summary. Table 3 shows various systems with their ranks based on ROUGE-2 and the average log-likelihood scores.

3.3 Correlation of ROUGE and log-likelihood scores

Tables 2 and 3 display log-likelihood scores of various systems in the descending order of log-likelihood scores along with their respective ROUGE-2 scores. We computed the pearson correlation coefficient (ρ) of 'ROUGE-2 and log-likelihood' and 'ROUGE-SU4 and log-likelihood'. This was computed for systems (ID: *1-32*) (*r1*) and for humans (ID: *A-J*) (*r2*) separately, and for both distributions.

For the binomial model, *r1* = -0.66 and *r2* = 0.39 was obtained. This clearly indicates that there is a strong negative correlation between likelihood of occurrence of a non-query-term and ROUGE-2 score. That is, a strong positive correlation between likelihood of occurrence of a query-term and ROUGE-2 score. Similarly, for human summarizers there is a weak negative correlation between likelihood of occurrence of a query-term and ROUGE-2 score. The same correlation analysis applies to ROUGE-SU4 scores: *r1* = -0.66 and *r2* = 0.38.

Similar analysis with the multinomial model have been reported in Tables 4 and 5. Tables 4 and 5 show the correlation among ROUGE-2 and log-likelihood scores for systems[2] and humans[3].

ρ	**ROUGE-2**	**ROUGE-SU4**
binomial	-0.66	-0.66
multinomial	-0.73	-0.73

Table 4: Correlation of ROUGE measures with log-likelihood scores for automated systems

ρ	**ROUGE-2**	**ROUGE-SU4**
binomial	0.39	0.38
multinomial	0.15	0.09

Table 5: Correlation of ROUGE measures with log-likelihood scores for humans

4 Conclusions and Discussion

Our results underscore the differences between human and machine generated summaries. Based on Summary Content Unit (SCU) level analysis of query-bias we argue that most systems are better at finding important sentences only from *query-biased* sentences. More importantly, we show that on an average, 76.67% of the sentences picked by any automated summarizer are *query-biased*. When asked to produce query-focused summaries, humans do not rely to the same extent on the repetition of query terms.

We further confirm based on the likelihood of emitting non query-biased sentence, that there is a strong (negative) correlation among systems' likelihood score and ROUGE score, which suggests that systems are trying to improve performance based on ROUGE metrics by being biased towards the *query terms*. On the other hand, humans do not rely on *query-bias*, though we do not have statistically significant evidence to suggest it. We have also speculated that the multinomial model helps in better capturing the variance across the systems since it distinguishes among *query-biased* sentences by quantifying the amount of query-bias.

From our point of view, most of the extractive summarization algorithms are formalized based on a bag-of-words query model. The innovation with individual approaches has been in formulating the actual algorithm on top of the query model. We speculate that the real difference in human summarizers and automated summarizers could be in the way a query (or relevance) is represented. Traditional query models from IR literature have been used in summarization research thus far, and though some previous work (Amini and Usunier, 2007) tries to address this issue using contextual query expansion, new models to represent the query is perhaps one way to induce topic-focus on the summary. IR-like query models, which are designed to handle 'short keyword queries', are perhaps not capable of handling 'an elaborate query' in case of summarization. Since the notion of *query-focus* is apparently missing in any or all of the algorithms, the future summarization algorithms must try to incorporate this while designing new algorithms.

[2] All the results in Table 4 are statistically significant with p-value ($p < 0.00004$, N=32)

[3] None of the results in Table 5 are statistically significant with p-value ($p > 0.265$, N=10)

Acknowledgements

We thank Dr Charles L A Clarke at the University of Waterloo for his deep reviews and discussions on earlier versions of the paper. We are also grateful to all the anonymous reviewers for their valuable comments.

References

Massih R. Amini and Nicolas Usunier. 2007. A contextual query expansion approach by term clustering for robust text summarization. In *the proceedings of Document Understanding Conference*.

John M. Conroy, Judith D. Schlesinger, Jade Goldstein, and Dianne P. O'leary. 2004. Left-brain/right-brain multi-document summarization. In *the proceedings of Document Understanding Conference (DUC) 2004*.

Terry Copeck, D Inkpen, Anna Kazantseva, A Kennedy, D Kipp, Vivi Nastase, and Stan Szpakowicz. 2006. Leveraging duc. In *proceedings of DUC 2006*.

Hoa Trang Dang. 2005. Overview of duc 2005. In *proceedings of Document Understanding Conference*.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *the proceedings of ACM SIGIR'95*, pages 68–73. ACM.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *the proceedings of ACL Workshop on Text Summarization Branches Out*. ACL.

H.P. Luhn. 1958. The automatic creation of literature abstracts. In *IBM Journal of Research and Development, Vol. 2, No. 2, pp. 159-165, April 1958.*

Daniel Marcu and Laurie Gerber. 2001. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proceedings of the NAACL-2001 Workshop on Automatic Summarization*.

Ani Nenkova, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: exploring the factors that influence summarization. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 573–580, New York, NY, USA. ACM.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. In *ACM Trans. Speech Lang. Process.*, volume 4, New York, NY, USA. ACM.

G.J. Rath, A. Resnick, and R. Savage. 1961. The formation of abstracts by the selection of sentences: Part 1: Sentence selection by man and machines. In *Journal of American Documentation.*, pages 139–208.

Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *the proceedings of IJCAI '07.*, pages 2862–2867. IJCAI.

Kristina Toutanova, Chris Brockett, Michael Gamon, Jagadeesh Jagarlamundi, Hisami Suzuki, and Lucy Vanderwende. 2007. The pythy summarization system: Microsoft research at duc 2007. In *the proceedings of Document Understanding Conference*.

Hans van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL 03 Text summarization workshop*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.