Bayesian Summarization at DUC and a Suggestion for Extrinsic Evaluation

Hal Daumé III and **Daniel Marcu**
Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
{hdaume,marcu}@isi.edu

Abstract

We describe our entry into the Document Understanding Conference competition for evaluating query-focused multi-document summarization systems. Our system is based on a Bayesian Query-Focused Summarization model, similar to the system we entered into the MSE competition. This paper begins by describing the (few) differences between our DUC system and our MSE system and describes our placement in the competition. The remainder of this paper argues in favor of performing *extrinsic* evaluation of summarization systems, and suggests a method for doing so.

1 Our DUC System

The system we entered into DUC this year is nearly identical to the system we entered into MSE a few months ago. Please refer to (Daumé III and Marcu, 2005) for more details on the system. In brief, the system is based on a Bayesian Query-Focused Summarization (BQFS) model (manuscript, unpublished). This model can be considered a black box that takes as input a collection of queries, a collection of documents and links that connect queries to relevant documents (relevance judgments). As output, the system will produce scores for each sentence in each document for its respective query. These scores are generalized distances, so that a sentence achieving a score of zero is best, and all other scores are strictly positive. Using these scores, we can easily rank sentences for extraction.

Once sentences have been scored by the BQFS method, we use the same discriminative training technique described in our MSE paper to perform the actual sentence selection. This aspect of the model learns weights for the following sentence-level features: similarity to previously extracted sentences via MMR (Carbonell and Goldstein, 1998), score according to the BQFS model, position score, a binary feature that detects quotes, the sentence length, the document length, the document similarity to the mean document, KL divergence between the sentence and the centroid document, the number of pronouns in this sentence, and the number of attribution verbs in this sentence. We use binary search on feature weights, one feature at a time, with ROUGE-BE as the objective function. Finally, like our MSE system, we perform weak sentence compression, by dropping constituents of various labels. Unlike MSE, this turned out to not be so useful for optimizing ROUGE-BE performance on development data. We believe this is because MSE aimed for 100 word summaries, while DUC aimed for 250 word summaries; we believe the former case to be more interesting.

2 DUC Results

We briefly describe three sets of the official DUC results. In Figure 1, we have graphed the official pyramid scores, both precision-based and recall-based; we have also presented the aggregated f-score. We have not included the human summaries in this graph, though they score better than any other

	Q1	Q2	Q3	Q4	Q5
Best Human	4.97	4.97	5.00	5.00	5.00
Best System	4.34	4.74	4.14	4.50	4.00
Med Human	4.84	4.91	4.94	4.91	4.81
Med System	3.92	4.46	2.98	3.20	2.10
Our System	2.60	4.04	3.12	3.08	1.88

Table 1: Linguistic quality scores for best and median systems and humans, as well as for our system. Scores are from 1 to 5; higher is better.

system. Our system, system 10, is the left-most bar in each section. We score second place for precision and first place (by a larger margin) for recall, given us the overall highest score. (These results are for the edited, full pyramids; the results for the raw full pyramids are similar, the results for the reduced pyramids are completely different for an unknown reason.)[1]

The second set of results we present are the human evaluation results, though not based on the pyramid methods. The first section in Figure 2 is a *responsiveness* score (how well did the summary answer the question) and the second section is a *linguistic quality* score (scaled up by 5 to fit in the same scale). The first ten columns in each section (highest columns) are the human summaries; our system (10) scored highest among systems for responsiveness and rather low (only beating four other systems) for linguistic quality (most likely due to the rather ad hoc sentence compression we performed).[2]

The final set of results we present are the automatic evaluation results, in Figure 3. In this figure, the left section is for ROUGE-SU4, while the right column is for ROUGE-2. Again, the first ten columns in each section are human results; our system (10) score sixth among systems for ROUGE-SU4 and third for ROUGE-2.

3 System Discussion

While we were pleased at our systems performance with respect to responsiveness, we were a bit surprised our system performed so poorly according to the linguistic quality evaluation. To further evaluate this, in Table 1, we have presented linguistic quality assessments for the four DUC questions for: the best human (for each question), the best system (for each question), the median human, the median system and our system. The questions are as follows:

Q1. Grammaticality: no datelines, system-internal formatting, capitalization errors or ungrammatical sentences.

Q2. Non-redundancy: no unnecessary repetition in the form of whole sentences, facts or noun phrases.

Q3. Referential clarity: no dangling references.

Q4. Focus: no sentences containing irrelevant information.

Q5. Structure and coherence: no disconnected facts; summary should build into a coherent body of information.

As we can see from Table 1, our system falls significantly below the median for grammaticality (Q1), and slightly below for non-redundancy (Q2), focus (Q4) and coherence (Q5). We are slightly above the median for referential clarity (Q3). Our relatively poor grammaticality performance is almost certainly due to the simplistic sentence compression we employ. Our slightly better performance on referential clarity most likely has to do with the fact that we explicitly include a feature in our model for looking at the presence of pronouns, and we have observed that after training the model, this feature receives a negative weight (sentences with many pronouns are disprefered).

Another observation one can draw from Table 1 is that the best system (and the median system) suffer most on Q3 and Q5: referential clarity and structure and coherence (the latter being the worst). This suggests that further research into these areas is needed for all systems, not just for our own.

4 Toward Extrinsic Evaluation

Evaluation of summarization systems is difficult. As a community, DUC has, over the past several years, employed a variety of manual and automatic measures of summary quality, including ROUGE (Lin and Hovy, 2003), SEE (Lin, 2001) and pyramid (Nenkova and Passonneau, 2004). Additionally, we

[1]Thanks to Liang Zhou for aggregating these results.
[2]Thanks to Guy Lapalme for aggregating these results.

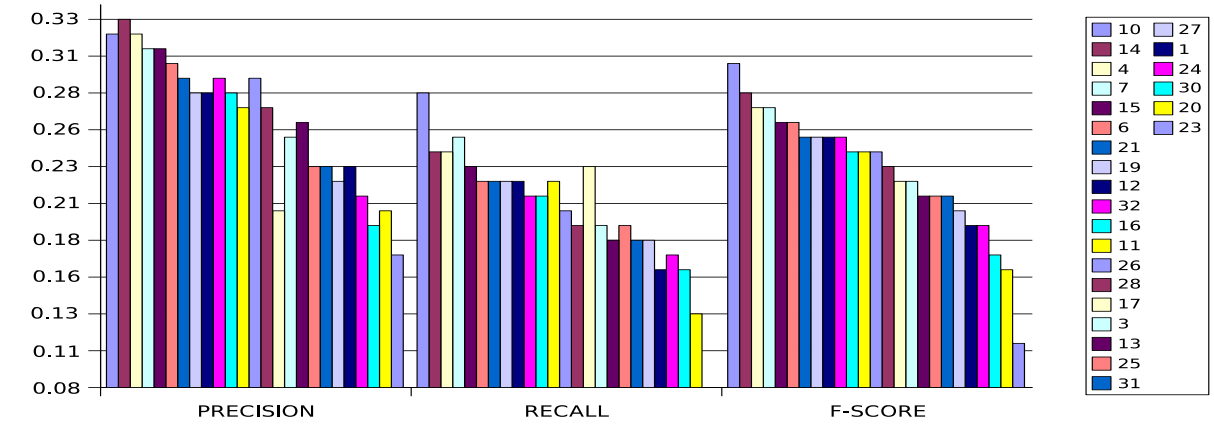

Figure 1: Official DUC pyramid results; we are system 10.

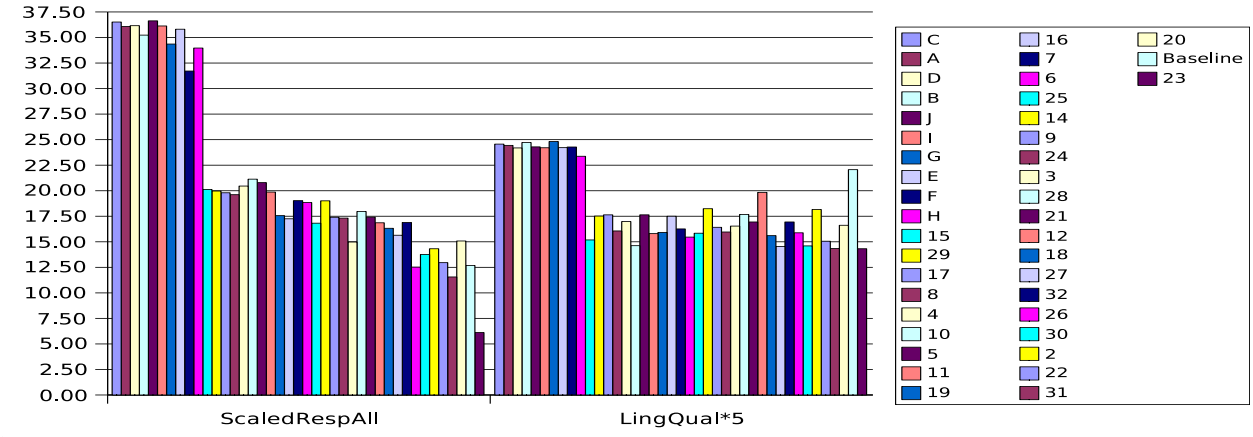

Figure 2: Official DUC human evaluation results; we are system 10.

have this year employed the linguistic quality questions (derived from SEE) and a responsiveness question (related to the fact that this year's summaries are query-focused). With the exception of ROUGE, all of these metrics are manually computed and are increasingly time consuming and expensive (pyramid, especially so). In this section, we argue in favor of moving away from these *intrinsic* evaluation metrics and toward *extrinsic* evaluation metrics. We will not discuss automatic metrics at all.

4.1 Why Extrinsic?

When one wishes to solve a real-world task, the world hands to us an evaluation metric. For instance, in information retrieval, a reasonable evaluation metric is: does the system improve my ability to find information quickly. Unfortunately, this metric is often too difficult or too time-consuming or too expensive to measure. When this is the case, one often seeks a surrogate measure that is expected to correlate well with the true metric. Again, in IR, measures such as mean average precision, or mean reciprocal rank, have been used to evaluate IR systems without having to set up large-scale experiments.

All of the manual DUC evaluation measures have been of the latter variety: we have separated the *task* (producing summaries that enable users to find information relevant to their individual needs) from the *measure* (how well does a system summary match what a human would do when asked to perform this task). By separating the task from the measure, we have been able to evaluate summarization systems without embedding them in a larger task.

Unfortunately, our current methods for evaluating

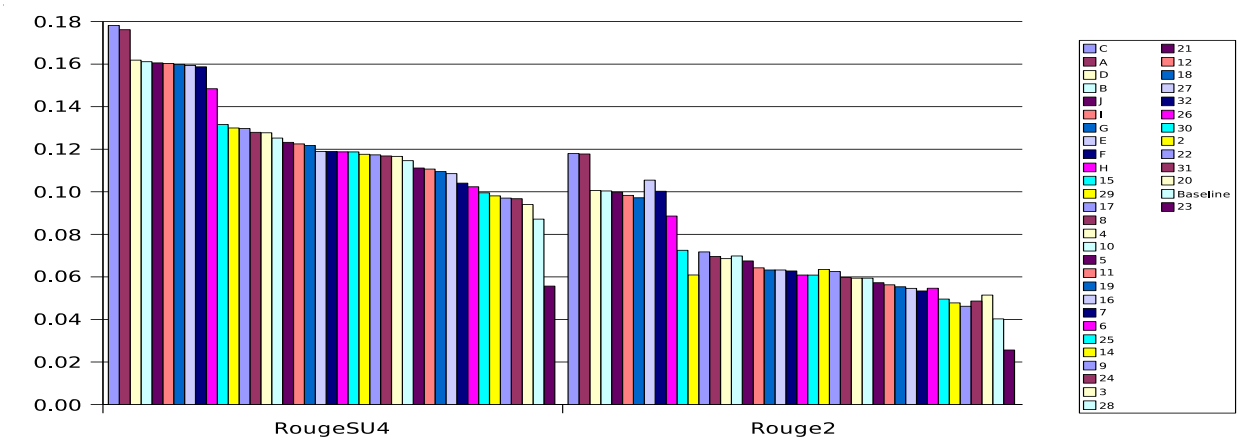

Figure 3: Official DUC automatic ROUGE results; we are system 10.

summarization systems intrinsically are time consuming and error prone. Given that we are already expending a huge amount of time doing system evaluation, we find it interesting to consider whether we might not be better off performing an *extrinsic* evaluation, instead. This presents its own problems, most importantly, how to select a task. We will shortly present one possible task, which we believe to be both realistic and easy to perform, but it is likely that having a small collection of maximally orthogonal tasks would be worthwhile. Moreover, it is important to keep in mind that when performing an intrinsic evaluation, as we have been doing, we are implicitly assuming that better performance on the intrinsic evaluation will lead to better performance on a task. However, at this point, we wonder what sort of task the current evaluation systems measure: Do high pyramid scores imply good performance for web search? For law search? For product comparison?

4.2 Proposal: Relevance Prediction

We propose the use of summaries in the context of a search application. This is inherently a query-focused single document task. The idea behind the relevance-prediction evaluation scheme is that a query-focused summary of a document should give sufficient information about the document to be able to judge relevance, without reading the full document.

The task is executed as follows: a large collection of documents is assembled, and a number of queries (e.g., 20) are assembled similarly to those for TREC (we ensure that there are *some* relevant documents in the set, but not too many). A baseline IR system is run against the collection and documents are judged by humans for relevance. For each query, a number of relevant documents and a number of non-relevant documents are set aside (perhaps 10-20 of each). A summarization system is then given the query and the documents and asked to produce a single-document summary of each (both for the relevant and the non-relevant documents). A human, different from the one used for the original document judgments, now reads each of the queries and summaries and based on only this information, attempts to judge whether the document is or is not relevant. A system that enables the judge to correctly predict relevance is better than one that does not.

This evaluation metric has several good points. First and most importantly, improved performance in this task clearly has applications to several real world tasks. Second, the more expensive annotation—the original relevance annotations—has already been done by NIST for the TREC evaluations; reading the summaries and judging relevance would be a reasonably efficient process, and not unlike the task we all perform daily when interacting with a search engine. Anyone could do such an evaluation; we would not need specially trained people (aside from being able to understand the queries). Finally, since the evaluation is based on the output of a real IR system, it is unlikely that simple tech-

niques such as pulling out the top ranked words, is not likely to perform well (because most IR engines use something like tf-idf for ranking, so the tf-idf scores of most top rated documents will be similar).

A potential disadvantage to this approach is that it might be advantageous for a system to attempt to preclassify each document as relevant or non-relevant, and build a bogus (or empty) summary for documents it deems irrelevant. In this case, the evaluation will measure something other than summary quality. We do not believe this to be a significant issue, so long as the system that performs the initial retrieval is state-of-the-art. In that case, improvements in classification performance beyond that done implicitly by the IR system is unlikely.

One possible controversial aspect of this evaluation is that it does not require any humans to write any summaries. This is good, because it significantly cuts down the amount of human effort needed to perform the evaluation. The only downside is that it does not provide the DUC community with additional training data for building systems.

4.3 Previous Experiments with this Metric

This metric has been proposed before (Dorr et al., 2004; Zajic et al., 2004; Dorr et al., 2005) to providing a litmus test for automatic evaluation metrics. The experiments we describe here can be seen as additional evidence that this metric is potentially useful, but looking at it from a different perspective. The systems we compare using this metric are also different from those compared previously.

4.4 Evaluation Experiments

We have performed some initial experiments validating the effectiveness and efficiency of this approach. We have used the TREC data sets and TREC relevance judgments as our data source. We used queries 204, 205, 207, 208, 210, 211, 212, 216, 217, 220, 221, 223, 225, 227, 228, 234, 235, 238, 239, 240, 242, 243 and 248 in the experiments. For each query, we took a collection of (on average) 16 documents, divided unevenly between known relevant documents and known irrelevant documents. The irrelevant documents are the highest-rated documents for a given query from the best performing system from the corresponding TREC evaluation.

Four summarization systems were run on these document/query pairs, aiming at a summary length of 40 words. The first, PREFIX, simply took the first 40 words. The second, TF-IDF, selected out the 40 words from a document with the highest TF-IDF scores and presented these as a bag of words. The third, KWIC, performed key-word extraction, similar to that done by major search engines. This system identified the locations of the query terms, and took windows of three words on each side until the 40 word limit was reached. The final system, KL, performed sentence extraction using KL-divergence between sentences and queries as the ranking function. Both the KL system and the KWIC system had access only to the short title section of the query.

System	P	R	F
KWIC	71.0	48.2	57.4
KL	65.6	47.7	55.2
PREFIX	67.6	36.3	47.3
TF-IDF	51.8	29.2	37.4

Table 2: Precision, recall and f-score of the filtering task for the four systems evaluated.

We have six human evaluators perform the evaluation. They were presented with the query (title, description and summary) and the summaries of the documents (both relevant and irrelevant). They were instructed to select those documents that they thought were likely to be relevant to the query. In total, we obtained results on 2772 document/query pairs. On average, it took annotators approximately 100 second to evaluate one query (corresponding to roughly 16 document/query pairs).

We evaluated each system to ascertain whether the humans using this system could correctly separate relevant from irrelevant documents, based only on the summaries. We computed this by looking at the precision, recall and f-score between the true relevance judgments and the relevance judgments based on the summaries alone. These results are shown in Table 2. As we can see from this table, the KWIC system performed best, followed by the sentence extraction KL system. The PREFIX system performed significantly worse than either the KWIC and KL systems, and the TF-IDF system performed significantly worse again. Interestingly, KWIC did not universally dominate KL: in 40% of

the queries, KL performed better. This suggests that a hybrid method (i.e., a sentence extraction system with compression capabilities) should be able to do better than either alone.

4.5 Agreement Experiments

In order to judge annotator agreement, we had multiple annotators judge the same queries. We computed the kappa score on these multiple annotations (in a pairwise fashion, 2 categories, 2 codes, 1260 data points) and achieved a score of $\kappa = 0.424$, which was not as high as we would have hoped. In order to better understand these numbers, we also computed the kappa values on a per-system basis (for instance, for a bad system, we would expect humans to have more difficulty deciding relevance and thus have lower agreement). These results are as follows: for KWIC (341 points), $\kappa = 0.513$; for KWIC (269 points), $\kappa = 0.495$; for PREFIX (348 points), $\kappa = 0.437$; for TF-IDF (302 points), $\kappa = 0.226$.

These kappa values, which top out at 0.513, are still lower than we would like. However, in post-evaluation discussion with evaluators, there arose several issues that would need to be sorted out before such an evaluation were used on a large scale, and which might serve to improve agreement. The most pertinent issue was: how strict should the annotation be. For instance, someone in desperate need of information, would eventually click every link until something is found. On the other hand, someone only mildly curious in a bit of information might only select two documents before giving up. Some effort would need to be put into making this explicit before deploying this evaluation metric. A second issue that came up was that the TF-IDF system was incredibly difficult to evaluate. This is also seen by the shockingly low (0.226) kappa values for this system. This system also took roughly 50% more time to evaluate than the others, since each word was high-content and out of context.

4.6 Discussion of Evaluation

In this section, we have argued in favor of moving toward extrinsic evaluation metrics, essentially because we are spending so much time and money on evaluation as is, it makes sense to consider evaluations that more closely measure performance for a particular task. We have suggested the use of the *relevance-prediction* metric, which we have found to not be time consuming and to provide at least an intuitive ranking of four basline systems on sample data. Moreover, the best system only achieves an f-score of 57.4, leaving significant room for improvement. Since this metric is inherently a single-document, query-focused summarization task, it might be worthwhile to investigate other tasks for which multidocument summarization is natural (for instance, product review summaries).

Acknowledgments

This work was supported by DARPA-ITO grant N66001-00-1-9814 and NSF grants IIS-0326276 and IIS-0097846.

References

Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336.

Hal Daumé III and Daniel Marcu. 2005. Bayesian multi-document summarization at MSE. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures*.

Bonnie Dorr, Christof Monz, Douglas Oard, Stacy President, David Zajic, and Richard Schwartz. 2004. Extrinsic evaluation of automatic metrics for summarization. Technical Report LAMP-TR-115, CAR-TR-999, CS-TR-4610, UMIACS-TR-2004-48, University of Maryland, College Park.

Bonnie Dorr, Christof Monz, Stacy President, and Richard Schwartz. 2005. A methodology for extrinsic evaluation of text summarization: Does ROUGE correlate? In *Proceedings of the Association for Computational Linguistics Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technology (NAACL/HLT)*, Edmonton, Canada, May 27 – June 1.

Chin-Yew Lin. 2001. SEE – Summary Evaluation Environment. `http://www.isi.edu/~cyl/SEE/`.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technology (NAACL/HLT)*, Boston, MA, USA, May.

David Zajic, Bonnie Dorr, Richard Schwartz, and Stacy President. 2004. Headline evaluation experiment results. Technical Report LAMP-TR-111, CS-TR-4573, UMIACS-TR-2004-18, University of Maryland, College Park.