DUC in Context

Paul Over {over@nist.gov}
Hoa Dang {hoa.dang@nist.gov}
Donna Harman {donna.harman@nist.gov}
Retrieval Group
Information Access Division
National Institute of Standards and Technology
Gaithersburg, MD 20899-8940, USA

Abstract

Recent years have seen increased interest in text summarization with emphasis on evaluation of prototype systems. Many factors can affect the design of such evaluations, requiring choices among competing alternatives. This paper examines several major themes running through three evaluations: SUMMAC, NTCIR, and DUC, with a concentration on DUC. The themes are extrinsic and intrinsic evaluation, evaluation procedures and methods, generic versus focused summaries, single- and multi-document summaries, length and compression issues, extracts versus abstracts, and issues with genre.

1 Introduction

Recent years have seen increased interest in text summarization with emphasis on evaluation of prototype systems [1]. Many factors can affect the design of such evaluations, requiring choices among competing alternatives. The realization of such designs seldom goes entirely as planned and the evaluations have complex effects on the researchers and their work.

What issues have the major evaluations addressed, what choices have they made and why, and what have been the consequences? This paper examines several major themes running through the Document Understanding Conference (DUC) evaluations (2001 - 2006) but also present in the Summarization Evaluation Conference (SUMMAC) and the National Institute for Informatics Test Collection for IR (NTCIR) systems workshops.

SUMMAC (Mani et al., 1999) was a large-scale evaluation of text summarization systems that took place in 1998 as part of the Defense Advanced Research Projects Administration (DARPA) TIPSTER program. There were 16 systems that took part, and two major summarization tasks that were evaluated in some manner. The Japanese NTCIR evaluations included summarization tasks in 2000, 2002, and 2004 with about 10 systems working on two different summarization tasks each year.

In 2000 a new summarization evaluation program was begun, again initially sponsored by DARPA; a group of expert summarization researchers contributed to a roadmap (Baldwin et al., 2000) that provided guidance for DUC, with a pilot run in 2000, and the first major evaluation in the fall of 2001. The roadmap called for evaluation of summaries of both single documents and sets of multiple documents, at specified levels of text compression. It suggested that the initial evaluation was to be intrinsic (direct evaluation), with extrinsic evaluation (looking at how the summary affects performance on a task) to be phased in over time, along with requirements of deeper text understanding techniques that can lead to more complicated summaries.

Over the course of its first six years DUC has examined automatic single- and multi-document summarization of newspaper/wire articles, with both generic tasks and various focused tasks. The results have been evaluated in terms of linguistic quality as well as their completeness with respect to content chosen by human summarizers (or in comparison with very simple automatic systems run at NIST to serve as baselines). Participation has grown from 15 research groups to over three dozen.

Table 1 gives a quick summary of the various tasks and evaluation methodologies that have been used in DUC in 2001-2006, and provides a chronological view of the DUC evaluations. This paper, however, examines DUC not chronologically, but in the context of evaluation issues and in the context of the state-of-the-art in automatic summarization. Seven different but interconnected themes are explored.

1. intrinsic versus extrinsic evaluation
2. generic versus focused summaries
3. single- and multi-document summaries
4. length and compression issues
5. extracts versus abstracts
6. issues with genre
7. the evolution of specific DUC evaluation procedures and methods

[1] Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

2 Intrinsic versus extrinsic evaluation

Two major types of evaluation have been used for testing summaries: *intrinsic* evaluation where the emphasis is on measuring the quality of the created summary directly, and *extrinsic* evaluation where the emphasis is on measuring how well the summary aids performance on a given task.

Extrinsic evaluation requires the selection of a task that could use summarization and then measuring the effect of using automatic summaries instead of the original text. Critical issues here are the task selection and the metrics for measurement. Tasks should be time-consuming enough that summaries could be useful, but also be sensitive enough to the quality of a summary that differences among a set of well-constructed summaries will show a difference in performance of the task.

SUMMAC used extrinsic evaluation for two different real-world tasks.[2] The first task was that of quickly processing a list of documents to find the relevant ones. Ideally a user could read summaries in order to judge relevance, and then only have to fully examine a few of the documents, thus saving a great deal of time. SUMMAC worked with 20 Text Retrieval Conference (TREC) topics, looking at the top 50 documents retrieved by one search engine and asking systems to provide summaries of each document, both at a fixed length (10% of the source) and at a varied length (the system's choice). The metrics were the time it took humans to process the list and the accuracy of the judgments made, with comparisons made against using full text, using a baseline of the first 10% of the document and using the results of the various automatic summarizers. Results of this part of the evaluation showed that judging only 10% of the document did indeed take only half the time, but accuracy (particularly recall) suffered significantly. There were no significant differences between the systems for the fixed length summaries, and minimal differences at the varying lengths, with the better systems having longer summaries.

The second real-world task was that of categorization, i.e., given a set of 1000 documents, can these be manually separated into 10 categories using summaries as opposed to using the full text of the documents. This task turned out to be difficult to operationalize; the dry run used categories that were too easy to separate and the final evaluation used categories that were too hard to separate. For the final run there was little difference in the time taken using full documents and using summaries, with a significant loss of accuracy. There were no significant differences between the systems, probably because the task was too hard.

In 2004 the relevance judgment task was tried again at the University of Maryland (Dorr et al., 2004), this time with shorter (75-word) summaries from 9 different sources including the original headline (all documents were news articles) and a 75-word manually constructed summary. Results from this experiment showed a major difference in time-on-task (summaries took about 33% of the time to process), with no significant difference in accuracy between the use of the full text and the use of human-generated summaries. However there were significant losses of accuracy even using the best automatic summaries, and there were no significant differences between performance of the various automatic systems.

It should be noted that there are two distinct reasons for having an extrinsic evaluation. The first is to show that the use of summaries could actually help in a real-world task. Experiments such as the one at the University of Maryland in 2004 showed a major time advantage for the task of relevance judgments, with no accuracy lost when using human summaries. They also showed that automatic summary creation was significantly less accurate, leading to the assumption that not only is more research needed, but that this task is a good one to model for further evaluation (extrinsic or intrinsic).

However, the second reason for extrinsic evaluation is to detect differences across automatic systems, and this is much more difficult. The Maryland experiment showed no significant differences between the performances of the various automatic systems, and these systems included state-of-the-art summarizers. Whether the task (and metrics) were not sensitive enough to separate the systems or whether the systems were simply not that different is not clear, but this experiment illustrates the problems of extrinsic evaluation methodology. The Japanese NTCIR evaluation also worked on the relevance judgment task in 2000 (NTCIR2, 2001), using a similar evaluation as SUMMAC. One of the participating groups (Nakao, 2001) examined in

[2] A third task, that of question answering, was tried on a small scale in SUMMAC, and readers are referred to the report (Mani et al., 1999) for more on this pilot evaluation.

more detail the problems of using this extrinsic evaluation to find differences between summarization systems.

Intrinsic evaluation measures the quality of a summary directly. This requires decisions on what is important to measure, what metrics should be used, and then how to operationalize the evaluation. Critical issues here are ensuring that the summary qualities being measured are truly important qualities, that the metrics being used are sensitive enough to measure those qualities, and that the evaluation itself does not introduce major interpretation problems such as may occur due to human variability.

DUC has mainly concentrated on intrinsic evaluation, attempting to run the evaluation at a large enough scale to allow significant differences between systems to be discovered, when they existed. Within DUC, intrinsic evaluation has comprised direct judgments of both linguistic well-formedness and the degree to which an automatically created summary expresses the same content as a manually created one (starting from the same set of documents to be summarized). Extrinsic evaluation has played a much smaller part; DUC has included some very limited (simulated) extrinsic evaluation, by measuring the *usefulness* of single-document summaries in 2003, and the *responsiveness* of multi-document summaries for the question-answering tasks in 2003-2006 (asking judges to rate how well a summary answered the questions). Section 8 details these evaluations.

NTCIR in 2000 used an intrinsic evaluation, looking both at extracts and at abstracts of single Japanese newspaper articles. For the extracts they measured the number of matches between sentences selected by humans as being important versus those selected by the automatic systems, and for abstracts they used professional newspaper captioners to rank the automatic summaries of the documents on a 4-point scale according to readability and coverage of important content. This task was extended in 2002 (NTCIR3, 2002) to include testing of multi-document summaries, using a similar evaluation. Additionally they tried a new evaluation method that measured the number of manual edits (inserts, deletes, and replacements) that would be needed to revise the summaries for content and readability. For 2004 (NTCIR4, 2004) they continued these tasks and evaluation methods, but also used a Japanese variation of the DUC linguistic quality questions (see section 8.1) and a SUMMAC-inspired pseudo question-answering evaluation.

It should be noted that intrinsic and extrinsic evaluation are parts of a wide spectrum of evaluations, starting with basic intrinsic evaluation, e.g., how readable is a summary, moving through more task-dependent intrinsic evaluation such as how much of the important content of a document was covered, and finally moving to extrinsic testing for a given task, either at the laboratory level or with real users at a job site. All parts of this spectrum are needed, as each part reveals different aspects of system behavior; but some parts are easier to operationalize than others.

3 Generic versus focused summaries

The history of summarization has concentrated on the production of generic summaries, that is, summaries that are produced with only minimal specification regarding their intended situation, audience, and use. The idea of producing automatic abstracts of single documents was the initial driver of research, and generic summarization has formed the bulk of research up until recently. It was therefore natural that the DUC roadmap called for generic summary evaluation rather than using focused summaries that respond to some specific purpose or information need.

However, focus in the summarization task has been a major issue in the DUC evaluations. Is focus needed, and if so, how does one guide systems to address it? Can systems step up to the challenge?

Focus can be reflected in all three classes of factors involved in summarization (Sparck-Jones, 1998, 2001): input source, intended purpose, and the form of the output summary. In DUC 2001 and 2002 the 60 document sets to be summarized were chosen to fit certain types (single event, multiple events of a single type, subject, biographical, opinion), which might vary not only in the kinds of information they contained but in how they were organized and according to other sub-genre characteristics. This was to give a minimum focus based on the input source. No type-related requirements were put on either the intended purpose or the output summary. The summaries were to be generic, intended for an educated adult newspaper reader with no specific use in mind other than saving time by reading a condensation that at some level covered all the input in the documents being summarized. The instructions given to the humans that created the

model (also known as "manual" or "gold standard") summaries reflected these assumptions.

Radical differences in approaches or results based on document type were not found in 2001 and 2002. There was, however, continuing discussion about whether generic summaries were in fact realistic targets - did anyone really want such a thing or are all summaries in fact focused in some way, e.g., by keywords. Additionally, participants were unhappy with generic summaries because it was believed they increased the inevitable differences between different humans' summaries of the same documents. Such differences placed a ceiling on the best score the evaluation could measure using only one manual reference summary for each document set.

As a result, and after having worked with generic summaries for two years, attempts were made to focus the summarization tasks in DUC 2003 - 2006. In DUC 2003 documents were chosen from multiple sources to provide multiple articles on the same topic from about the same time period. Documents were chosen around topic detection and tracking (TDT) events/topics. In addition, two new sorts of information about intended purpose were introduced: viewpoints (short statements of specific interest within a broader topic) and question topics from TREC's Novelty track. In 2004 focus was provided by TDT events and 50 questions of the form "Who is X?", where X was a person. In 2005 and 2006 the system task modeled complex question answering, i.e., given a set of 25-50 documents per question, create a brief, fluent, well-organized answer to a question that cannot be answered just with a name, date, quantity, etc. (Dang, 2005).

For the most part systems have used the focusing material (topics, question, viewpoint statement) to select sentences from the text to be summarized. In some cases question answering techniques were applied (Lacatusu, Parker, & Harabagiu, 2003), (Blair-Goldensohn et al., 2004). In 2005 systems treated the complex question answering task as a passage retrieval task and ranked sentences according to their relevance to the topic. A significant minority of systems first decomposed the topic narrative into a set of simpler questions, and then extracted sentences to answer each question (Dang, 2005).

After considerable work on generic summaries, it is appropriate that DUC's emphasis should shift to focused summaries, and question-based summarization meshes nicely with current interest in question answering. The details of task definition for question-based summarization, indeed for all types of summarization, critically depend on a better understanding of how such summaries are used in real work situations. But note that there are needs for summaries at many points along the scale from minimally to maximally specified. Therefore there is a need to evaluate generic as well as focused summaries. Effective summarization technologies should be able to produce useful output with varying degrees of user/task context and do so flexibly based on user input/history.

4 Single- and multi-document summaries

DUC has addressed both single-document summarization and summarization of a set of documents on the same topic. The roadmap called for summarization of single documents — the traditional target of summarization systems. But the task of creating generic summaries of news articles (often by largely extractive means) turned out to be much less interesting than expected. Simple "take the lead sentence/paragraph" baselines could achieve very good results in news — the challenges of single-document summarization in other genres and for specific purposes await examination.

Summarization of multiple documents on a topic is also a very natural task and one that the roadmap identified as needing attention. Summarization systems function as part of a larger workflow and may well sit downstream from systems which filter and cluster documents on topics of standing interest from a much larger stream or collection of documents. The NewBlaster system (McKeown et al., 2002) from Columbia University and the WebInEssence system (Radev, Blair-Golddensohn, Zhang, & Raghavan, 2001) from the University of Michigan are several examples of operational systems using multi-document summarization.

Additionally using sets of documents can aid automatic summarization by providing information about what facts are important (and therefore repeated across documents). But having multiple documents may make even extraction more difficult since there are more choices to make and a greater likelihood of creating disfluencies. While from the beginning DUC systems have scored lower when summarizing multiple documents than when summarizing single ones, the scores are not dramatically different as can be seen

in Figures 1 and 2. As with single-document summaries, new genres and intended uses will present new challenges for multi-document summarization.

5 Length and compression issues

The length of the output summary was initially felt to be an important characteristic for users to be able to control and a key factor in system effectiveness to be investigated. In 2001 and 2002, target multi-document summary lengths of 50, 100, 200, and 400 words were set. While scores generally dropped as the target size decreased, results showed little difference in the relative performance of systems based on target size. Table 1 shows the various lengths that have been used for single- and multiple-document summaries.

For single-document summaries, it was decided to focus on a size of particular interest — very short (<= 75 bytes) summaries — in 2003 and 2004. Summaries of this length are familiar to many users of web search engines and allow them to choose among search results without looking at each in detail. In creating these very short summaries, groups applied sophisticated means such as linguistically motivated sentence compression with statistically selected topic terms (Zajic, Dorr, & Schwartz, 2004) and syntactic document analysis to produce pagerank-scored logical form graphs, parts of which could be extracted and merged to generate summaries (Vanderwende, Banko, & Menezes, 2004).

For the summaries of multiple documents, the lengths have also been selected as appropriate for the task, especially for the focused summary tasks started in 2003. The selected lengths needed to be short enough to encourage work beyond simple extracts, but long enough to allow sufficient levels of detail to be included in the summaries.

Different summary lengths result in different criteria for judging the overall quality of the summary. Very short summaries (<= 75 bytes) can have a non-standard grammar and still be considered "good" useful summaries (c.f. headlines). Longer summaries on the order of 200 words, however, require additional linguistic structure in order to be interpretable by humans; poor referential clarity and discourse structure, for example, can produce a confusing or, worse, misleading representation of the content of the original text.

DUC systems have handled almost all degrees of compression with few special means. But this is likely another consequence of choosing news articles as the input genre and generic summaries as the intended output. DUC to-date cannot shed any light on the compression for other genres as input and/or other intended purposes for the output.

6 Extracts and abstracts

The DUC organizers expected that participants, coming mostly from the natural language processing community, would quickly move beyond extraction to address the problems of deeper analysis of material to be summarized and to emphasize the synthesis of summaries. This has not generally happened except in the creation of very short summaries. It can be instructive to see what occurred and consider in retrospect why.

The approaches used in DUC have been largely extractive, i.e., they have been mainly concerned with selecting appropriate sentences from the material to be summarized and ordering them, largely unchanged, to create the output summary. Looking at the 2005 task for example, most participants treated it as passage retrieval. Sentences were ranked according to relevance to the topic. The most relevant sentences were then selected for inclusion in the summary while minimizing redundancy within the summary, up to the maximum 250-word allowance.

This was true despite the fact that NIST took some care in how the model (manual) summaries were created so as not to encourage cutting and pasting from the documents to be summarized. Summarizers were encouraged to express, at some level, all the content of documents to be summarized, thus encouraging, it was thought, use of generalizing language. And in fact one study found that no more than 55% of the vocabulary items found in a given model summary occur in the corresponding source document(s) (Copeck & Szpakowicz, 2004).

A significant minority of systems did first decompose the topic narrative into a set of simpler questions, and then extracted sentences to answer each subquestion (Jagadeesh, Pingali, & Varma, 2005). Systems differed in the approach taken to compute relevance and redundancy, using similarity metrics ranging from simple term frequency to semantic graph matching (Melli et al., 2005). In order to include more relevant information in the summary, systems attempted within-sentence compression by removing phrases such as parentheticals and relative clauses (Farzindar, Rozon, & Lapalme, 2005). Cross-sentence dependencies had to be handled, including anaphora. Strategies for dealing with pronouns that occurred in relevant sentences included co-reference resolution (Schilder, McCulloh, Thomson McInnes, & Zhou, 2005), including the previous sentence for additional context (Lacatusu, Hickl, Aarseth, & Taylor, 2005), or simply excluding all sentences containing any pronouns.

But most systems made no attempt to reword the extracted sentences to improve the readability of the final summary. Although some systems like Columbia University's (Blair-Goldensohn, 2005) grouped related sentences together to improve cohesion, the most common heuristic to improve readability was simply to order the extracted sentences by document date and position in the document. The LAKE05 system (D'Avanzo & Magnini, 2005) achieved high readability scores by choosing a single representative document and extracting sentences in the order of appearance in that document. This approach is similar to the baseline summarizer and produces summaries that are more fluent than those constructed from multiple document.

Beyond the inertia within research groups that are already experienced in extractive techniques, it seems likely that once again the use of news articles, especially at less than extreme compression, may have made development of abstractive techniques unnecessary. In addition, it was not clear that the tasks being modeled required summaries with substantial rewording. DUC has cast doubt on the assumption that abstracts are the end goal for most summaries. What real task, if any, will make the case for abstracts over extracts?

7 Issues with genre

Newspaper articles are part of the vast open source literature of interest to many people including the US intelligence community. Such material has been the basis for research in information retrieval (TREC), information extraction (MUC), topic detection and tracking (TDT), and summarization (SUMMAC). In large part the choice of newspaper articles followed from their availability and the fact that research groups had already worked on this genre.

The use of newspaper and newswire text as material to summarize was decisive for the DUC evaluations in a number of ways. The pyramidal structure of newspaper articles meant that simple baseline systems creating summaries from the first sentence(s) in a article or even a set of articles were difficult to beat.

Even once the effects of the choice were clear, the community has been slow to change direction and work on a new genre and new tasks. In 2003 work was begun on an updated summarization roadmap (Sparck-Jones et al., 2004) and the results of this effort and various other possibilities were discussed at DUC 2004. Difficulties of various sorts with all the proposals for new input genres and/or new tasks made it hard to find a consensus for more than incremental change. Possibilities and concerns included:

- speech - would errorfulness be fatal?
- fiction - could systems handle large amounts of data? what is the real task to model?
- email - where can one get the data (can privacy concerns be overcome)?
- emergency situation data - too much novelty of input genre, output requirements, etc.?
- blogs - possible, interesting, but issues little understood

The decision was made in 2004 (for DUC 2005) to keep working on the news text but to make the task more difficult (and realistic). The task of complex question answering, using multiple documents as input was selected, and this task was run both in 2005 and 2006 (the short evaluation cycle in 2006 suggested task repetition). DUC 2006 included both a simple (manual) exercise in timeline generation and another open

discussion of future tasks. The timeline generation task, while shown to be feasible, did not generate much interest. However, the task of evolving summaries (originally part of the roadmap suggestions in 2000), was considered reasonable and will be done on a pilot basis in DUC 2007. Additionally there was discussion of the use of blogs as input data and this is likely to be a new direction in the future.

8 The evolution of DUC evaluation procedures and metrics

From the beginning, the DUC evaluations have tried to evaluate automatically produced summaries along two dimensions: their linguistic well-formedness and the degree to which their content agrees with human-created summaries of the same material (coverage). These two dimensions are not independent since extreme lack of well-formedness can affect the ability to judge content overlap. There has been significant evolution in the evaluation of both dimensions, especially in coverage.

8.1 Linguistic quality

A set of linguistic quality questions was designed primarily to give detailed diagnostic information to system designers on questions of current interest to them. For the first DUC (2001) there were three quality questions:

- Grammaticality: Do the sentences, clauses, phrases, etc., follow the basic rules of English?
- Cohesion: Do the sentences fit in as they should with the surrounding sentences?
- Organization: Is the content expressed and arranged in an effective manner?

Human judges (assessors) found it difficult to distinguish between cohesion and organization and found the questions did not apply to very small summaries. As most of the approaches extracted whole sentences from the material to be summarized, the grammaticality scores were measuring human, not system, summarization performance.

The set of linguistic quality questions was expanded to 12 in 2002 and then revised into a smaller set of 7 in 2004 and then to 5 in 2005 and 2006:

- Grammaticality: The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.
- Non-redundancy: There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or repeated use of a noun or noun phrase when a pronoun would suffice.
- Referential Clarity: It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference should be unclear if an entity is referenced but its identity or relation to the story is unclear.
- Focus: The summary should have a focus; sentences should only contain information that is related to the rest of the summary.
- Structure and Coherence: The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

In DUC 2006, as in 2005, all summarizers generally performed well on the first two linguistic qualities. Participating systems scored higher on focus in 2006 than in 2005, with the best systems achieving scores comparable to humans. As a group, systems' performance remained unchanged on referential clarity and structure and coherence, though the best systems did come close to human performance on these qualities. See Figure 3 for a graphic representation of DUC 2006 linguistic quality results.

To the extent the questions address intrasentential properties, they are of little use when systems, as has frequently been the case, take sentences as-is from the documents to be summarized and incorporate them in the automatic summary. As systems increasingly try to incorporate abstractive techniques demanded by users and specific work contexts, diagnostic measurement of linguistic well-formedness will become more important. DUC has built and tested a foundation for further work.

8.2 Content coverage

Summarization research aims to create systems that can summarize like humans. Therefore content coverage in DUC has been understood to be the degree to which one summary (automatically created) conveys the same information as another (manually created) — both summaries starting from the same documents and assumptions of purpose. Ideally content coverage would be evaluated automatically in terms of the meaning of the summaries, independent of the superficial means of expression, but automatic identification and comparison of the propositional content units are unsolved problems. As a result DUC has worked on approximating the ideal along several tracks. DUC started with units that are superficial but can be identified automatically and with manual comparison of those units, but has also encouraged and used work in manual identification of meaningful units, automatic identification of meaningful units, and automatic matching.

SEE

From 2001 - 2004, the judging of content coverage was carried out at a very detailed level in order that it also could provide low-level diagnostic information. The Summary Evaluation Environment (SEE) (Lin, 2001) was developed for this purpose. Each summary to be evaluated (peer) and each reference summary (model) was divided into roughly clausal units. Peer summaries were automatically divided into units based on elementary discourse units (Marcu, 2000) while the model summaries were automatically divided into sentences. Then each unit in each peer was judged for its coverage of a corresponding model, with each even partially covered unit of the peer being identified as such.

Figure 4 shows an example coverage evaluation session using SEE. In the screen snapshot the evaluator is comparing the second model unit (upper right) to all of the peer units (upper left) and has identified the third peer unit as overlapping in meaning with the second model unit. The lower half of the screen shows that the evaluator has decided that all the overlapping peer units (perhaps there is only one), taken together, express about 40% of the content of the second model unit.

The procedure was time-consuming and consequently each peer could be judged against only one model. It treated all units of the models as though they were of equal value even though this is known not to be true. While the procedure produced detailed comparisons, researchers felt the clausal units were in fact too large since each could express several basic facts.

Since humans differ significantly in what material they choose in creating a summary, diagnostic information and scoring was limited by and dependent on one human summarizer's choices. Few if any systems ever reached the level of human-human agreement, so there was always room to improve. Overall coverage evaluation results were stable (Harman & Over, 2004); i.e., despite large variations in the human-generated model summaries and large variations in the manual judgments of single-model coverage, the ranking of systems remained comparatively constant when averaged over dozens of document sets, dozens of peer summaries, and 10 or so judges. However, the evaluation, as designed, did not provide system developers with as much training information as they wanted in order to compare to the variety of human summarizers' (or users') viewpoints.

ROUGE

Researchers wanted an automatic means of scoring coverage. They also desired a means (automatic or not) that would take into account the range of variation exhibited in human summaries, reflect the differences in relative importance among the piece of information included, and use a unit smaller than the clausal ones used in SEE. A software package, ROUGE, was developed to meet some of these needs (Lin, 2004).

ROUGE implements a family of automatic evaluation methods focused on recall and based on counting the number of text units in common between the peer summary and one or more model summaries. The text units can be N-grams, word sequence, word pairs, etc. Extensive experiments with ROUGE have demonstrated reasonable correlation with manual coverage judgments that makes it useful in system development via hill-climbing (Lin, 2004). But ROUGE's treatment of multi-word expressions and function words is not ideal (Hovy, Lin, Zhou, & Fukumoto, 2005).

Basic Elements

The Basics Elements (BE) package for summary scoring uses chunks of text generally larger than the tokens on which ROUGE was based but smaller than the clausal structures used in SEE. After some experimentation a basic element was defined as a head-modifier-relation triple, which includes the head of a major syntactic constituent (noun, verb, adjective, or adverbial phrase) and a relation between a head-BE and a single dependent (Hovy et al., 2005).

Pyramid

Another approach, similar to the Basic Elements, but currently requiring significant manual effort, is the Pyramid Method (Nenkova & Passonneau, 2004), in which humans identify the basic units of meaning — summary content units (SCUs) — in a set of model summaries. The more model summaries an SCU occurs in, the greater the weight it is given in scoring a peer summary that contains it. This method takes into account the variability of human summarizers, weights the semantic units based on their frequency across models, and provides detailed diagnostic information about why a given summary scores as it does. The use of humans in creating the pyramids and scoring summaries against the pyramids allows for great flexibility in recognizing the same basic semantic propositions even if they appear in very different forms. Attempts at automating the scoring have been made (Harnly, Nenkova, Passonneau, & Rambow, 2005).

Usefulness and Responsiveness

In DUC 2003, two simulated extrinsic evaluation measures were introduced on a small scale: *usefulness* of very short single-document summaries (<= 10 words), and *responsiveness* of multi-document question-focused summaries. Usefulness was inspired by extrinsic evaluations based on relevance judgments. For usefulness, the assessor sees a document and all summaries of that document. Working under the assumption that the document is one that they should read, the assessor grades each summary on a 5-point scale according to how useful it would be in getting them to choose the document. It was found that usefulness tracked average SEE coverage for these very short summaries.

Responsiveness is a NTCIR-inspired measure that ranks the summaries on a 5-point scale, indicating how well the summary satisfies a given information need. As with usefulness, responsiveness does not involve comparison of the peer summaries to any models, but peers may be compared to each other. Responsiveness generally tracked SEE coverage in 2003.

In DUC 2004 the use of SEE had to be restricted because of time pressure, but some type of manual evaluation test was needed. One of the DUC 2004 tasks required summaries in response to questions, so responsiveness was repeated on a larger scale for this task. Responsiveness was shown in 2004 to correlate well with results from SEE and became the manual evaluation performed at NIST in later DUCs (Over, 2004).

The automatic evaluation methods (ROUGE-2, ROUGE-SU4, and BE-HM) all correlate about equally well with manual content responsiveness. The correlations as shown in Table 2 suggest any of the methods

should be useful for system development if developers proceed carefully, knowing that automatic means depend more brittlely on word choice and sequence. Also, when trying to detect real differences in system performance, the automatic methods may have less discriminative power. For example, as Figure 5 shows, while the manual content responsiveness metric (x-axis) clearly separates the models from the peers, the difference between the models and the peers is quite small based on the automatic metric (y-axis).

9 Conclusions and prospects

Over the years, datasets, tasks, and systems have changed, as well as metrics and evaluation procedures. Nevertheless, DUC coverage results have been similar in the following ways:

- most manual summaries are clearly better than most automatic summaries
- most automatic summaries do not differ significantly
- automatic summaries at the extremes usually differ significantly
- automatic summaries seldom performed better than simple baselines based on the structure of news articles

Manual comparison of summaries generally supports the validity of the above findings.

While some uses of summaries may not require well-formed sentences, all require good coverage of the content to be summarized. System developers need a variety of content coverage measurement tools that range from quick automatic means for initial repeated testing of new ideas many times per day, to more costly but more informative evaluations (including human judgments) that may only occur annually and are applied only to very promising techniques. DUC has been a forum for developing and refining new content evaluation metrics, procedures, and tools at several points along this spectrum.

Every evaluation is the result of choices among alternatives and so no single evaluation type can serve all the valid purposes. SUMMAC, NTCIR, and DUC have explored many of the possible types of evaluation and some of the main combinations of factors that define the space of possible summaries. DUC has provided a framework for considerable work on how to measure linguistic quality and content coverage (automatically and manually). It has explored generic summaries and ways of focusing them, multi- and single-document summaries, and varying degrees of compression (especially in very short summaries) — all this for one genre: news articles.

In the future DUC needs to move not only beyond news genre but into new summarization tasks that are inspired by real uses of summaries. Research in summarization is expanding, driven by a growing information overload, and therefore evaluations must evolve to serve more diverse needs. Some of these needs come from uses of summaries in specific domains, such as medicine or law, but basic summarization research is also required in order to fully understand the relationship of summarization techniques to the factors of input sources, output requirements, and intended uses for these summaries.

References

Baldwin, B., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., McKeown, K., Mittal, V., Moens, M., Radev, D., Jones, K. S., Sundheim, B., Teufel, S., Weischedel, R., & White, M. (2000). *An Evaluation Road Map for Summarization Research.* url:duc.nist.gov/roadmapping.html.

Blair-Goldensohn, S. (2005). *From Definitions to Complex Topics: Columbia University at DUC 2005.* url:duc.nist.gov/pubs/2005papers/columbiau.sasha.pdf.

Blair-Goldensohn, S., Evans, D., Hatzivassiloglou, V., McKeown, K., Nenkova, A., Passonneau, B., Schiffman, B., Schlaikjer, A., Siddharthan, A., & Siegelman, S. (2004). *Columbia University at DUC 2004.* url:duc.nist.gov/pubs/2004papers/columbia2.nenkova.ps.

Copeck, T., & Szpakowicz, S. (2004). *Vocabulary Agreement Among Model Summaries and Source Documents.* url:duc.nist.gov/pubs/2004papers/uottawa.copeck.final.pdf.

Dang, H. T. (2005). *Overview of DUC 2005.* url:duc.nist.gov/pubs/2005papers/OVERVIEW05.pdf.

D'Avanzo, E., & Magnini, B. (2005). *A Keyphrase-Based Approach to Summarization: the LAKE System at DUC-2005.* url:duc.nist.gov/pubs/2005papers/itc-irst.ernesto.pdf.

Dorr, B. J., Monz, C., Oard, D., President, S., Zajic, D., & Schartz, R. (2004). *Extrinsic Evaluation of Automatic Metrics for Summarization.* University of Maryland, College Park, MD, LAMP-TR-115, CAR-TR-999, CS-TR-4610, UMIACS-TR-2004.

Farzindar, A., Rozon, F., & Lapalme, G. (2005). *CATS A Topic-Oriented Multi-Document Summarization System at DUC 2005.* url:duc.nist.gov/pubs/2005papers/umontreal.lapalme.pdf.

Harman, D., & Over, P. (2004). The Effects of Human Variation in DUC Summarization Evaluation. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out* (p. 10-17). Barcelona, Spain.

Harnly, A., Nenkova, A., Passonneau, R., & Rambow, O. (2005). Automation of summary evaluation by the pyramid method. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2005).* Borovets, Bulgaria.

Hovy, E., Lin, C.-Y., Zhou, L., & Fukumoto, J. (2005). *Basic Elements.* url:hayden.isi.edu/BE/.

Jagadeesh, J., Pingali, P., & Varma, V. (2005). *A Relevance-Based Language Modeling Approach to DUC 2005.* url:duc.nist.gov/pubs/2005papers/iiit.jagadeesh.pdf.

Lacatusu, F., Hickl, A., Aarseth, P., & Taylor, L. (2005). *Lite-GISTexter at DUC 2005.* url:duc.nist.gov/pubs/2005papers/lcc.finley.pdf.

Lacatusu, V. F., Parker, P., & Harabagiu, S. M. (2003). *Lite-GISTexter:Generating Short Summaries with Minimal Resources.* url:duc.nist.gov/pubs/2003papers/lcc.ps.

Lin, C.-Y. (2001). *Summary Evaluation Environment (SEE).* url:hayden.isi.edu/SEE/.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out* (p. 74-81). Barcelona, Spain.

Mani, I., Firmin, T., House, D., Klein, G., Sundheim, B., & Hirschman, L. (1999). The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of the EACL'99: Ninth Conference of the European Chapter of the Association for Compuational Linguistics* (p. 77-85). Bergen, Norway.

Marcu, D. (2000). *The Theory and Practice of Discourse Parsing and Summarization.* MIT Press., Cambridge, Massachusetts.

McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J., Nenkova, A., Sable, C., Schiffman, B., & Sigelman, S. (2002). Tracking and Summarising News on a Daily Basis with Columbia's Newsblaster. In *Proceedings of the Second International Conference on Human Language Technology Research* (p. 280-285).

Melli, G., Wang, Y., Liu, Y., Kashani, M. M., Shi, Z., Gu, B., Sarkar, A., & Popowich, F. (2005). *Description of SQUASH, the SFU Question Answering Summary Handler for the DUC-2005 Summarization Task.* url:duc.nist.gov/pubs/2005papers/simonfraseru.sarkar.pdf.

Nakao, Y. (2001). How Small Distinction among Summaries can the Evaluation Method Identify. In *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization* (p. 235-241). Tokyo, Japan.

Nenkova, A., & Passonneau, R. (2004). Evaluating Content Selection in Summarization: The Pyramid Method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (p. 145-152). Boston, MA.

Over, P. (2004). *An Introduction to DUC 2004 - Intrisic Evaluation of Generic News Text Summarization Systems.* url:duc.nist.gov/pubs/2004slides/duc2004.intro.pdf.

Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies: Information Retrieval, Question-Answering and Summarization. (2004). url:research.nii.ac.jp/ntcir/workshop/onlineproceedings4/index.html.

Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese and Japanese Text Retrieval and Text Summarization. (2001). url:research.nii.ac.jp/ntcir/workshop/onlineproceedings2/index.html.

Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question-Answering. (2002). url:research.nii.ac.jp/ntcir/workshop/onlineproceedings3/index.html.

Radev, D., Blair-Golddensohn, S., Zhang, Z., & Raghavan, R. (2001). Newsinessence: A System for Domain-Independent, Real-Time Clustering and Multi-Document Summarization. In *Proceedings of the First International Conference on Human Language Technology Research* (p. 274-277).

Schilder, F., McCulloh, A., Thomson McInnes, B., & Zhou, A. (2005). *TLR at DUC: Tree Similarity.* url:duc.nist.gov/pubs/2005papers/thomson-lr.schilder.pdf.

Sparck-Jones, K. (1998). Advances in Automatic Text Smmarization. In I. Mani & M. T. Maybury (Eds.), (pp. 1–12). MIT Press., Cambridge, Massachusetts.

Sparck-Jones, K. (2001). *Factorial Summary Evaluation.* url:duc.nist.gov/pubs/2001papers/cambridge2.pdf.

Sparck-Jones, K., Halteren, H. van, Moens, M.-F., Lapalme, G., Radev, D., Dorr, B., Over, P., Hovy, E., McKeown, K., & Harman, D. (2004). *DUC Roadmap 2005-2007.* url:duc.nist.gov/RM0507/rm.html.

Vanderwende, L., Banko, M., & Menezes, A. (2004). *Event-Centric Summary Generation.* url:duc.nist.gov/pubs/2004papers/microsoft.banko.pdf.

Zajic, D., Dorr, B., & Schwartz, R. (2004). *BBN/UMD at DUC 2004: Topiary.* url:duc.nist.gov/pubs/2004papers/umaryland.zajic.pdf.

	2001	2002	2003	2004	2005	2006
Generic summaries (*length)*						
single document						
(*100)*	****	****				
(*10)*			****	****		
multiple document						
(*50, 100, 200, 400*)	****					
(*10, 50, 100, 200*)		****				
extracts (*200, 400*)		****				
Focused summaries (*length)*						
multiple document						
viewpoint (*100)*			****			
question/topic (*100)*			****			
event (*100)*			****	****		
'who is' question (*100)*				****		
complex question (*250)*					****	****
Corpus Size (doc sets x docs/set)	60x10	60x10	60x10 30x25	100x10	50x32	50x25
Manual Evaluation	SEE	SEE	SEE Responsiveness Usefulness	SEE Responsiveness	Pyramid Responsiveness	Pyramid Responsiveness
Automatic Evaluation				ROUGE	ROUGE/BE	ROUGE/BE

Table 1: Tasks and evaluation methodologies for DUC 2001-2006. Summary length is measured in number of words.

Metric	Spearman	Pearson
ROUGE-2	0.767	0.836 [0.725, 1.000]
ROUGE-SU4	0.790	0.850 [0.746, 1.000]
BE-HM	0.797	0.782 [0.641, 1.000]

Table 2: DUC 2006 correlation between average content responsiveness and average ROUGE-2/ROUGE-SU4/BE-HM recall over all automatic peers.

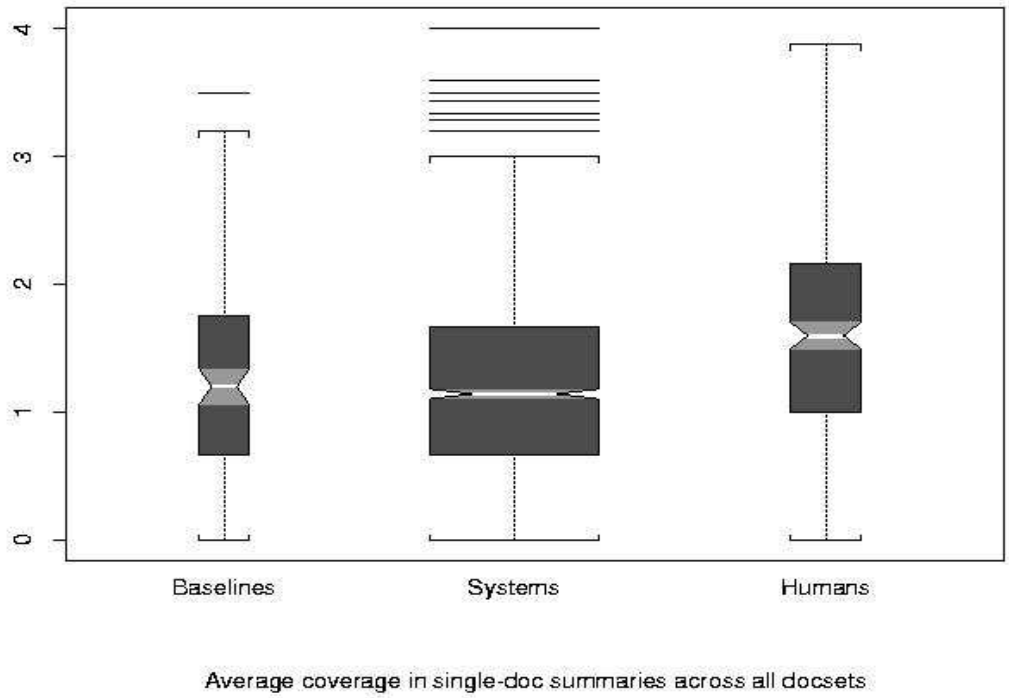

Figure 1: DUC 2001 coverage for single-document summaries (Model facts covered by the peer: 0 = none, 1 = hardly any, 2 = some, 3 = most, 4 = all)

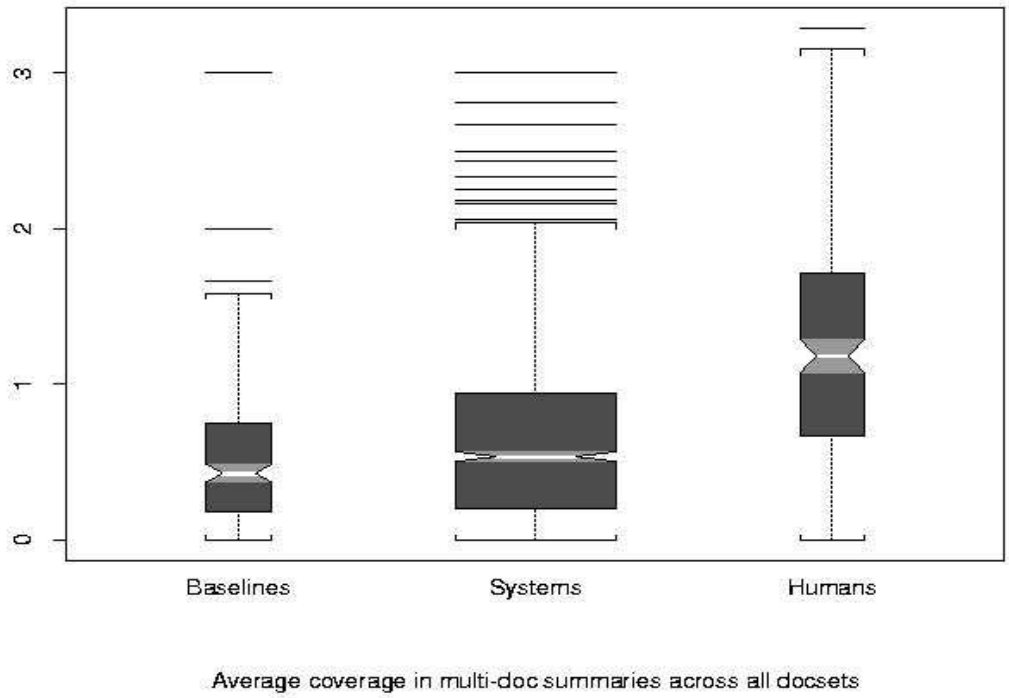

Figure 2: DUC 2001 coverage for multi-document summaries (Model facts covered by the peer: 0 = none, 1 = hardly any, 2 = some, 3 = most, 4 = all)

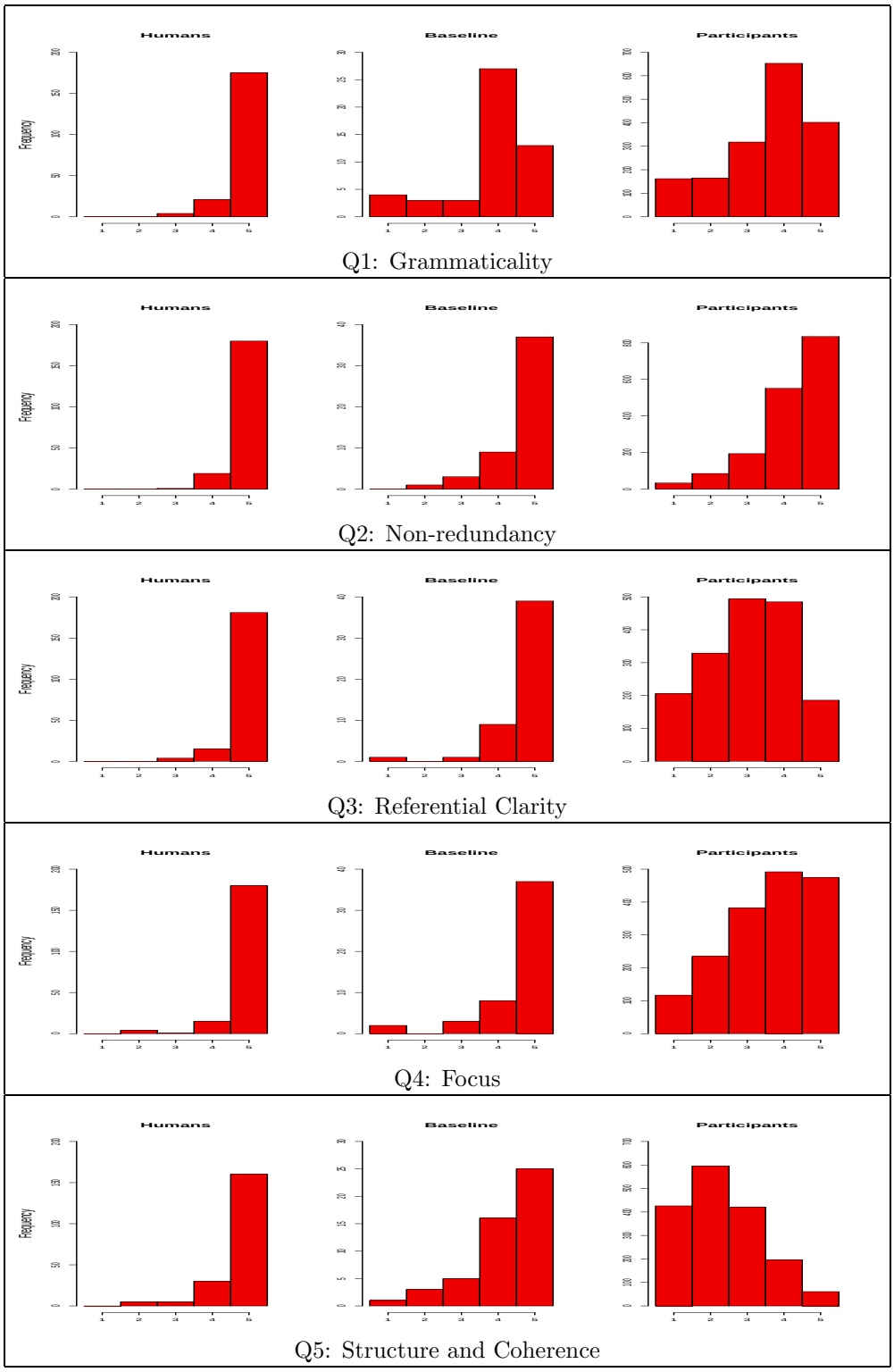

Figure 3: DUC 2006 frequency of scores (5 = best) for each linguistic quality, broken down by source of summary: Humans(left), Baseline(middle), Participants(right)

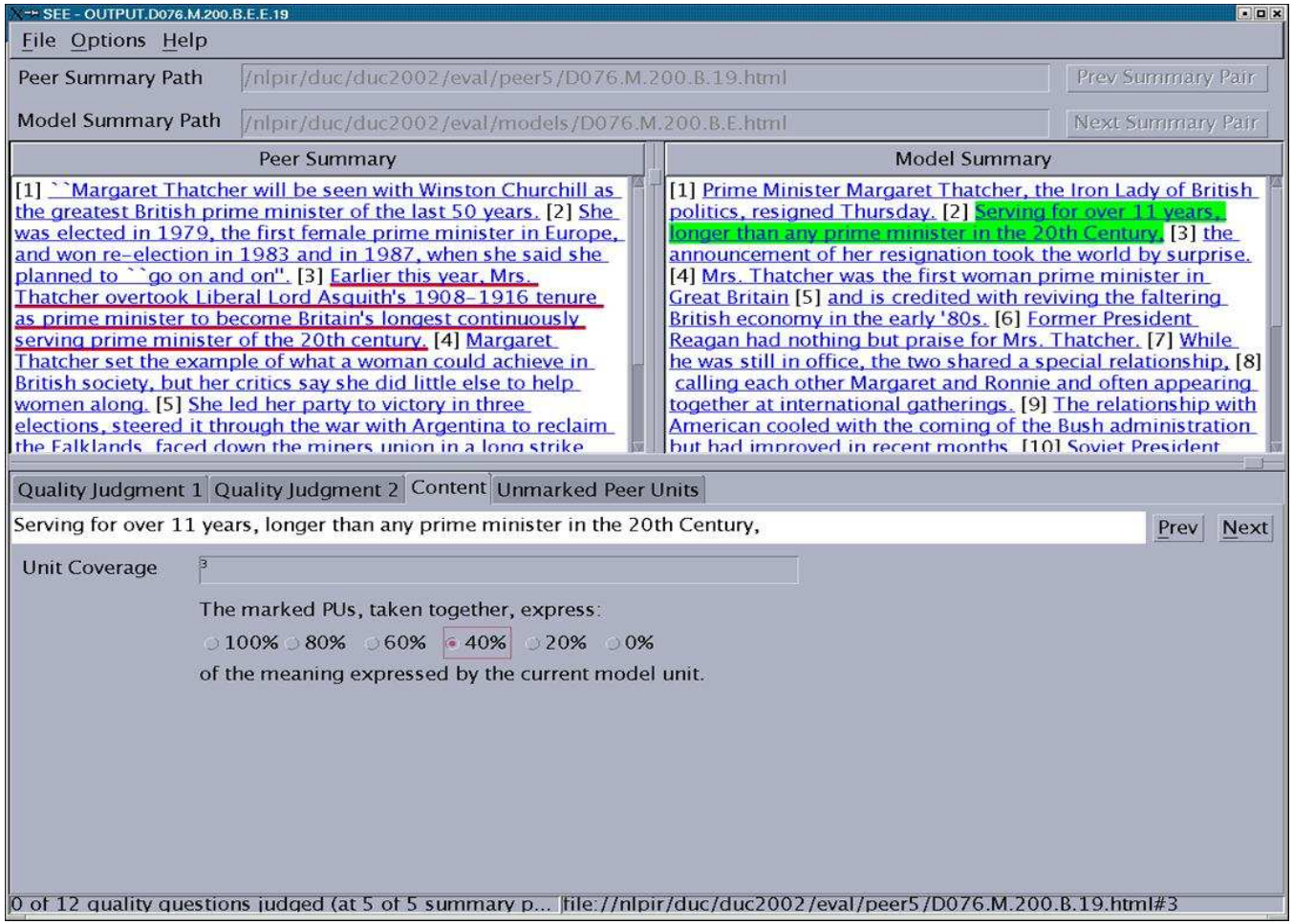

Figure 4: DUC 2002 Summary Evaluation Environment (SEE)

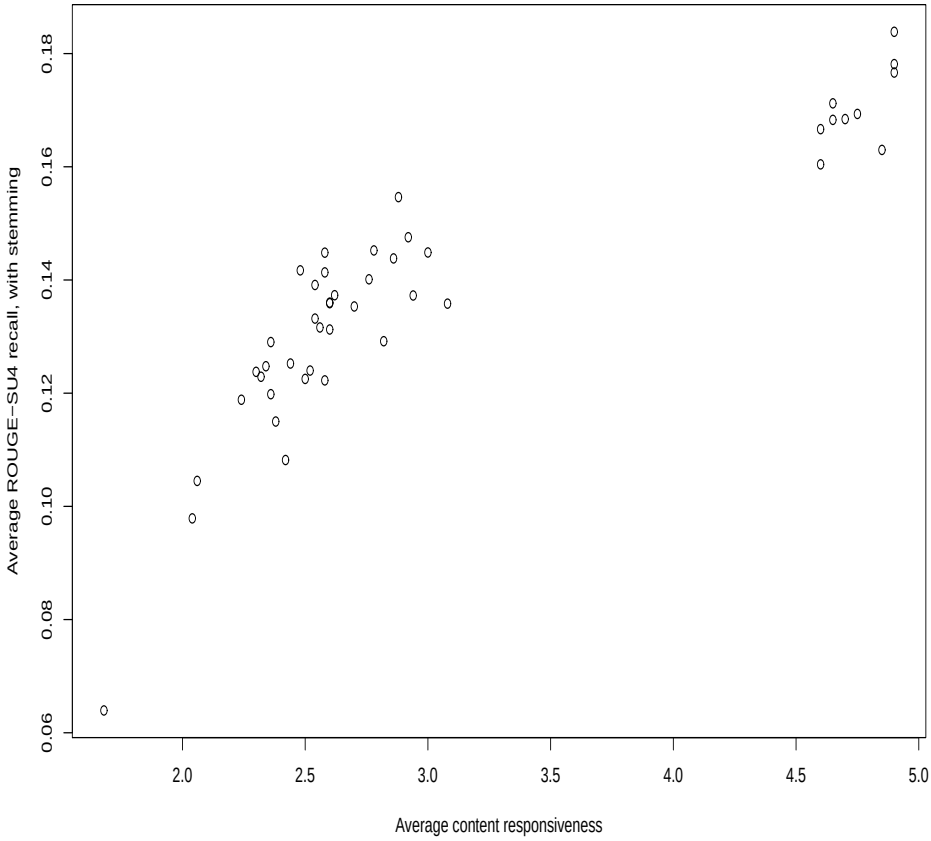

Figure 5: DUC 2006 average content responsiveness vs. average ROUGE-SU4 recall with stemming