From Definitions to Complex Topics: Columbia University at DUC 2005

Sasha Blair-Goldensohn
Department of Computer Science
Columbia University

October 3, 2005

Abstract

We describe our work in adapting a system designed to answer definitional and biographical questions (i.e. "What/Who is X?") to the topic-focused summarization task of DUC 2005. We present a system overview, focusing on newly developed aspects. We evaluate our performance, examining example output and analyzing our scores on the various metrics provided by NIST. We conclude with general observations about this year's task and the direction of future work at Columbia.

1 Introduction

The DUC 2005 task presented unique challenges. The document sets were larger and more heterogenous than in previous years. There was relatively little training data with which to prepare our systems. But above all, the biggest challenge were the long-form topics around which we were to create summaries: they were lengthy, complex, and often called for a good deal of inference and interpretation.

Nonetheless, we were hopeful that we could adapt our DefScriber question answering (QA) system [Blair-Goldensohn et al., 2004b] to this task. We were motivated in part by the success we had had in adapting this system from primarily definitional QA to produce biographical summaries in DUC 2004, where it tied for the top ROUGE scores [Blair-Goldensohn et al., 2004a]. However, we were not entirely certain the approach would adapt as well this time; for biographical summaries, we could essentially apply the definitional approach to "define" a person, wherea topics like those in Tables 1 and 2, there is not a clear single entity to define (although the topic title is sometimes a rough analog). Moreover, the extended topic descriptions often contained several nested questions and were quite challenging for a system to parse.

In order to handle this new task, we made several adaptations to DefScriber's design, mainly in the question parsing and passage retrieval modules. We then grafted these modified parts onto the robust, non-definition-specific parts of DefScriber. The base DefScriber system is overviewed in Section 2, with the adaptations detailed in Section 3.

Still, the complex topics were so daunting that we were initially unsure if our efforts to parse them would be in vain. To informally test this, we conducted an informal side-by-side evaluation of the adapted DefScriber and several competitive summarization systems under development at Columbia (these other systems effectively ignored the question statements and produced a summary based solely on document set content). In these informal comparisons, the adapted DefScriber seemed more effective than the pure summarization approaches, and thus we used it for our entry.

As we observe in Section 4, our successful results on the DUC 2005 test set validate this decision. We first examine an example output of the system. Then we then proceed with a formal analysis of the quantitative results, observing that our system achieved strong results in many of the metrics evaluated, and was one of the top perform-

Title VW/GM Industrial Espionage **Question** Explain the industrial espionage case involving VW and GM. Identify the issues, charges, people, and government involvement. Report the progress and resolution of the case. Include any other relevant factors or effects of the case on the industry.
Title Fertile Fields **Question** Discuss making and using compost for gardening. Include different types of compost, their uses, origins and benefi ts.

Table 1: Some example topics from the DUC 2005 test set (d311i, d694j).

ers in terms of average ranking across metrics.

2 System Overview

DefScriber was initially developed for definitional question answering as part of Columbia's work in the AQUAINT (Advanced Question Answering for Intelligence) program, and the core system was more recently adapted to add capability for the biographical summaries task in DUC 2004 [Blair-Goldensohn et al., 2004a]. (When we refer to DefScriber in this paper, we are referring to the core system which uses a set of definition-focused methods to create definitions of objects and concepts, as well as biographical summaries, i.e. the system as it existed before any modifications for the DUC 2005 task; we will call the adapted system for DUC 2005 DefScriber-A.) In this section, we will give a very brief overview of that original DefScriber system. We detail the adapted DefScriber-A system in the next section.

DefScriber's approach relies on a combination of goal-driven and data-driven techniques. The data-driven techniques shape answer content in a bottom-up manner, according to themes found in the data, using statistical techniques including centroid-based similarity [Radev et al., 2000] and clustering [Hovy and Lin, 1997]. The goal-driven techniques apply a top-down method, using a set of *definitional predicates* to identify types of information ideally suited for inclusion in a definition, such as hierarchical information (i.e., "X is a kind of Y distinguished by Z.").

The base DefScriber system is described in detail in [Blair-Goldensohn et al., 2004b]; following is brief a description of its processing pipeline:

1. **Identify relevant sentences** which contain information pertinent to the target individual or term (i.e. the X in the "Who/What is X?" question).

2. **Incrementally cluster extracted sentences** using a cosine distance metric, weighting with a combination of collection and local word-frequency IDF features.

3. **Select sentences for output summary** using a fitness function which maximizes inclusion of core definitional predicates, coverage of the highest-ranking clusters, and answer cohesion.

4. **Apply reference rewriting techniques** to extracted sentences to improve readability of summary, using an auxiliary system developed at Columbia [Nenkova and McKeown, 2003] and initially integrated as part of DUC 2004.

3 Adaptations for DUC 2005 Task

The key changes made for the DefScriber-A system adapted for the DUC 2005 task were in relevant-passage selection, or Step 1 in the pipeline described in the previous section.

The main criterion for determining sentence relevance in the original DefScriber is the concentration of words from the target name or term, i.e. the X in the "Who/What is X?" question. In that setting, question parsing simply amounts to taking all non-stopwords in the term. Sentences containing or nearby these terms are extremely likely to be classified as relevant. (The exact function we use for determining relevance also gives a light weight to some other features such as sentence length and position.)

However, for the DUC 2005 task, the complexity of the topic statements posed a significant challenge in terms of question parsing. Not only did we have a much more complex question to deal with, but also little training data around which to design and evaluate relevant-sentence detection algorithms. Given these limitations, we combined several robust techniques which we believed would perform acceptably on the unknown but sure-to-be-challenging questions in the test set.

1. **Term frequency-based weighting** Given the lengthy topic statements, we needed to have some way of filtering the more and less relevant terms from the topic statement ("topic terms"). Our approach was to assign each topic term a weighting proportional to its IDF as calculated over a large news corpus.

Title Threat to Wildlife by Poachers **Question** Where have poachers endangered wildlife, what wildlife has been endangered and what steps have been taken to prevent poaching?	If African elephants are to be saved, the economic return on elephant farming must be increased, rather than lowered, perhaps by granting export quotas to countries willing to invest in keeping the poachers out. The area on either side of the river teems with elephants and other game, now threatened by poachers. Kenya banned big game hunting in 1977 and this year said poachers would be shot on sight. Offi cials from Kenya's Wildlife Service, who have won plaudits worldwide for their anti-poaching efforts, say they need freedom to cross borders when pursuing poachers and smugglers. Tourists pay millions of dollars a year to come and see Africa's wildlife – and smugglers pay millions more in strictly illegal purchases of ivory and rhino horn from poachers. Until recently, rural communities were not allowed to make any use of wildlife on their lands - poaching was common, either for food or to stop animals destroying crops and endangering people. The number of poached carcasses of elephants and black rhinos in Luwangwa fell by 90 per cent between 1985 and 1987. Poaching has wiped out all but - at an optimistic estimate - 500 of Zimbabwe's rhinos; four years ago there were more than 2,000. Three have been shot already, and even more have been killed in Zimbabwe, the only other country with a shoot-to-kill policy toward poachers. Euan Anderson, a Zimbabwean vet, believes that since the dehorning programme started in Zimbabwe, four to fi ve dehorned rhinos have been killed by poachers.

Table 2: An example DUC 2005 topic (d407b) and DefScriber-A's answer.

2. **Topic structure** We further adjusted the topic term weights with the simple heuristic of giving terms in the title double the weight of terms in the extended question/topic body. However, we were unable to make use of the "granularity" setting given with the topic statements.

3. **Stemming** In order to maximize coverage of relevant terms when measuring overlap of topic terms and document sentences, we used Porter stemming and matched over word stems.

4. **Nearby-sentences** We informally experimented with several schemes for including the content of nearby sentences in the determination of a given sentence's relevance. These experiments indicated that a window of two sentences on either side was helpful, with the highest weight given to the immediately preceding sentence.

Using these techniques, we implemented an algorithm for determining on a per-sentence basis which sentences in the document set were relevant to a given topic statement. The algorithm made two passes over each document, on the first pass assigning relevance scores to each sentence based on overlap with topic terms (using weighting as explained above). In the second pass, these scores were adjusted using the first-pass scores of nearby sentences, and sentences scoring above a certain cutoff were judged relevant (additional sentences would be kept if less than 30 sentences were above the cutoff score).

In addition to the changes for relevant-sentence selection, we also made a change to Step 3 of the pipeline described in the previous section, i.e. the step where the output is created by selecting and ordering some number of the relevant sentences. The change here involved disabling the use of the top-down strategy which attempts to place sentences expressing "is-a" type information about the term/individual being described at the start of an answer. The reason for disabling this technique is that it assumes a certain model, i.e. that a single entity is being discussed. Given the topics which we saw in the training set, it seemed that this was not likely to be the case for most topics, and that following this heuristic was likely to decrease the relevance of our answers.

4 Results Analysis

We first performed an informal analysis of our system's answers, and found that in many cases the DefScriber-A system was successful in presenting a summary of the topic that was largely relevant to the question. An example question/answer pair is shown in Table 2. We can see that on this question our main modifications in DefScriber-A to identify topic-relevant sentences are successful: the output clearly covers the various aspects of the topic statement. In addition, we can see that the base techniques inherited from DefScriber are also effective in this setting, for example avoiding redundancy by use of clustering, and employing cohesion measures that put related parts of the answer together (e.g., grouping together sentences about Kenya and Zimbabwe).

In addition to this informal examination of our results, we performed various statistical tests to determine the significance of the quantitative results distributed by NIST. (In order to simplify the process of analysis and ranking, we excluded the scores for the human-produced summaries from the remainder of tests described here. We

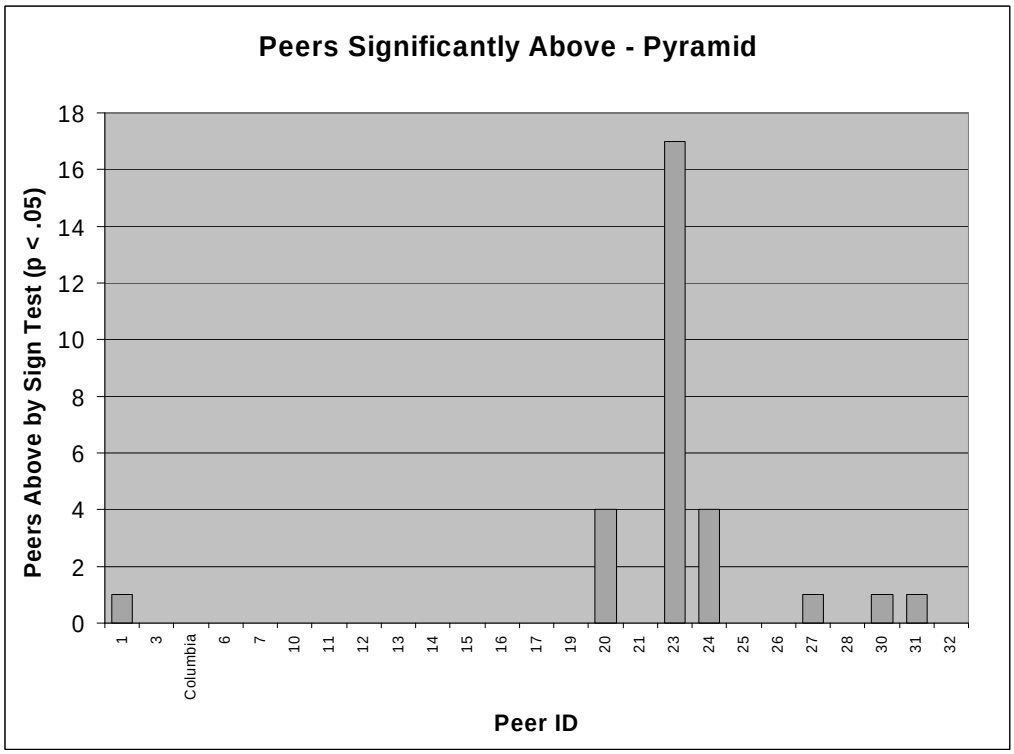

Figure 1: Systems-above rank for modified pyramid scores.

did this because we were primarily concerned with comparing systems to other systems, and because the results for all metrics clearly had all humans outperforming all systems.)

The first group of tests we carried out were two-way ANOVA analyses, to determine whether scores on the various metrics depended significantly (a) the system used and/or (b) the particular document set. The basic questions being answered by this test were (a) whether some systems have different levels of performance than others and (b) whether some document sets have different levels of difficulty than others. We did these tests to determine whether ranking of systems would be informative; this would only be the case if the choice of system indeed did have a significant effect on performance (the inspiration for this approach came from [Nenkova, 2005]).

All of the ANOVAs we performed showed significant effects from both factorscwith $p \leq 0.05$. This was true for all measures, including the linguistic quality (LQ) questions, responsiveness question, ROUGE recall scores (only ROUGE-2, L, and SU4 were analyzed), and modified pyramid scores (we used the scores from the processed_pans.txt file, taking the score for a given docset/system as the mean of the two annotator scores, as advised in the email from the pyramid organizers at Columbia).[1]

Given that the ANOVAs showed a significant effect, we proceeded to carry out rank tests to determine whether or not any individual comparisons between system rank reached the level of significant difference. We carried out both the Wilcoxon Matched-Pairs Signed Ranks test and the Sign test, and observed similar results on both; we report results on the Sign test only.[2]

Our results are summarized in Figures 1-5. We highlight the following observations:

1. **Pyramid** The Sign tests finds many significant dif-

[1]However, this tells us only that some systems and docsets had signifi cantly different means than other systems and docsets; in order to get a sense of how many individual system pairs had signifi cantly different means, we performed several one-way ANOVA experiments with Bonnferoni error correction to account for the multiple comparisons. Running this test on responsiveness, ROUGE-SU4 recall, and LQ question 1, we found that the proportion of system pairs with signifi cant differences at $p \leq 0.05$ by this test was 0.10, 0.24 and 0.17, respectively. Thus, it is possible that some of the rank differences we report according to the Sign test may occur between systems whose mean base scores were not signifi cantly different.

[2]On consideration, we report the results of these rank tests in terms of the number of peers ranked above (rather than below) a given system, since we believe it is more interesting to know how close to the best performers a given peer is, rather than how many lower performing systems are below it. Note that this means that *lower* numbers are better, since the best systems have no peer ranked above them.

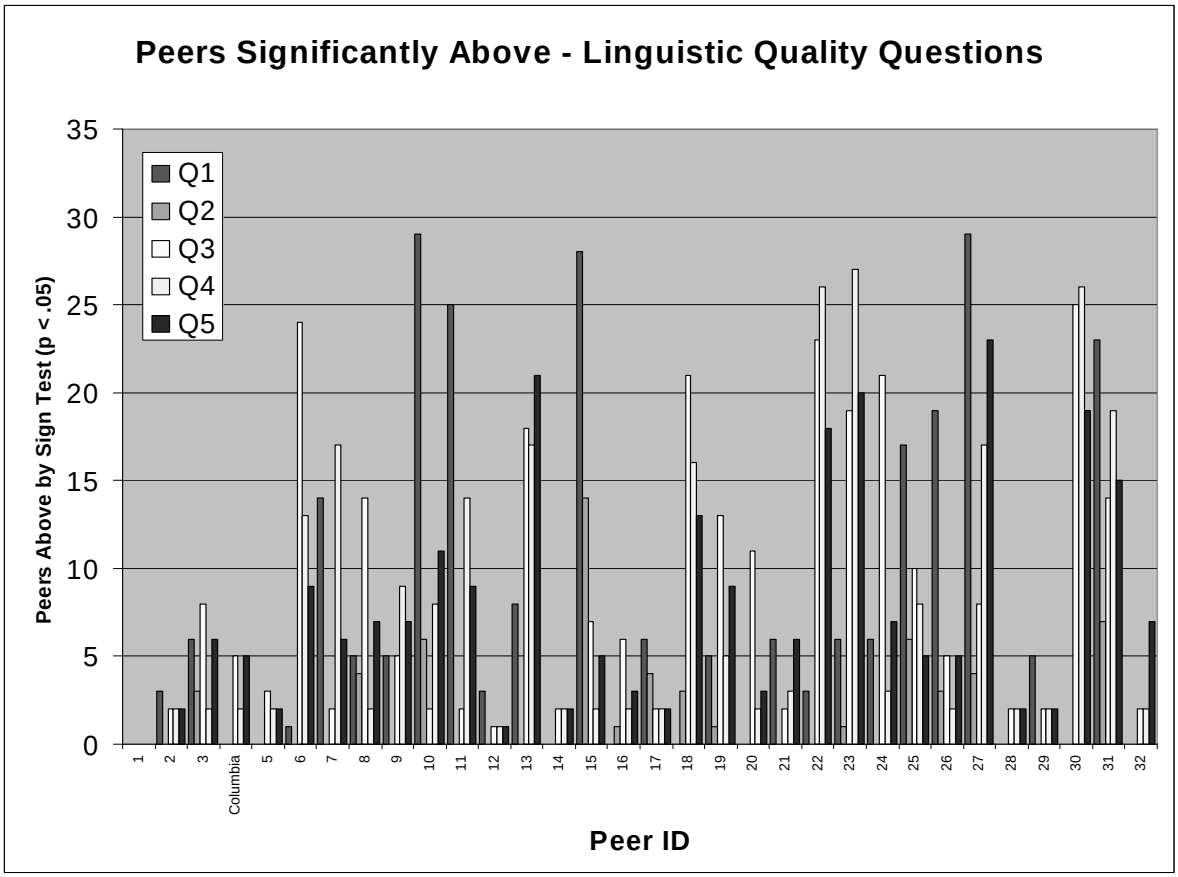

Figure 2: Systems-above rank for individual linguistic quality questions.

ferences in rank for most metrics. However, there were many fewer for pyramid than other measures, as shown in Figure 1. This is likely because of the smaller number of samples. Because of this, and because not all systems participated in the evaluation, we do not use the pyramid scores in the combined overall measures presented below.

Columbia ties with 17 of the 25 peers evaluated by pyramid for the best systems-above rank here (no systems ranked significantly higher).

2. **Linguistic Quality** Figure 2 shows the performance on the various linguistic quality (LQ) questions. Given the difficulty of reading each of the five questions in a single chart, we considered the possibility of taking the mean systems-above rank across the five questions on the theory that they all measure related aspects, i.e. matters of presentation as opposed to content. To justify this statistically, we performed Spearman rank correlation across all pairs of LQ rank measures, and found that they are all significantly correlated with $p \leq 0.05$, with the exception of the pairs (q1,q3) and (q2,q3), which both had a weaker, but still positive, correlation. Given this strong correlation, we use the mean systems-above rank across the LQ questions as shown in part of Figure 4

   Columbia ties with one other system as the 9th best in mean systems-above across LQ questions, with a mean of 2.4 peers ranked higher. Note also that the baseline system was by far the best performer on the LQ questions. This is not surprising given that the baseline was simply the first 250 words of a presumably well-written article.

3. **Responsiveness** Mean systems-above rank for the responsiveness question is shown as part of Figure 4. Columbia ties with 10 other peers for the best systems-above rank, with no systems ranked significantly higher.

4. **ROUGE** Figure 3 shows systems-above ranking for the systems across the ROUGE-2, L and SU4 recall measures. (These ranks are based on scores which

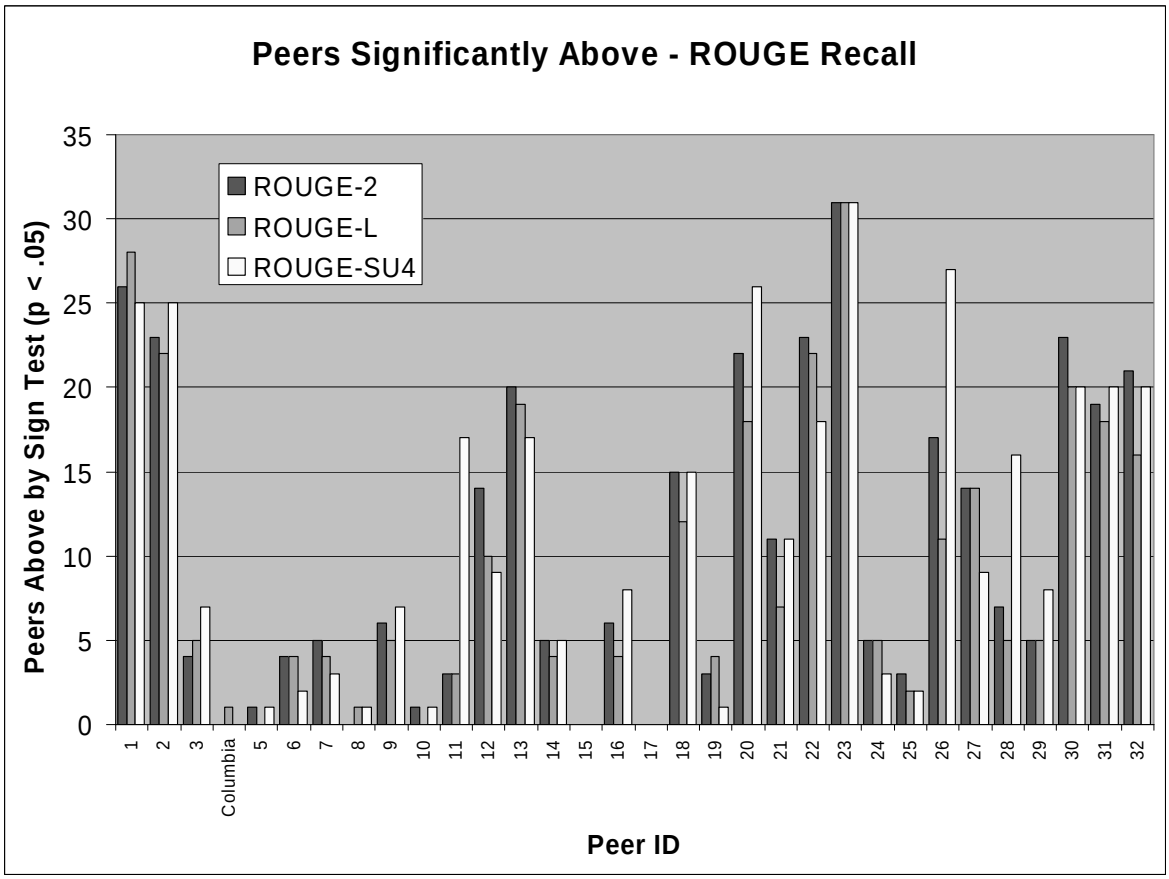

Figure 3: Systems-above rank across three ROUGE recall measures

take into account the issues mentioned in communication from NIST regarding the different number of models across different docsets, by ranking within docsets only for the base data.) These numbers are significantly correlated by Spearman, although ROUGE-L tends to produce slightly more significant differences in rank. However, for the score presented as part of Figure 4, we take the mean systems-above rank for ROUGE-2 and SU4, as these are the official NIST measures.

Columbia ties with three other system as 3rd best in mean systems-above across ROUGE-2 and SU4 recall, with a mean of 0.5 peers ranked higher.

5. **Overall** To get an overall picture of which systems performed consistently well across the various metrics, we combined the three measures displayed in Figure 4, namely mean systems-above rank for ROUGE-2 and SU4 recall, mean systems-above rank for LQ questions 1-5, and systems-above rank for the responsiveness. We then take the mean of these three measures, with the result shown in Figure 5. Motivations for using this particular combination include: (1) It combines measures of presentation and content, but more weighted toward content (which seems fair since getting good presentation alone is fairly trivial as the baseline showed) (2) It combines automatic and manual scores (3) It uses scores where there was a significant level of difference found between systems, and where all systems were rated.

   In the combined measure, Columbia is 2nd best, slightly behind peer 5 and slightly ahead of peer 17.

5 Conclusions and Future Work

Our participation in DUC 2005 was an excellent opportunity to evaluate the flexibility and extensibility of our question-answering work on a new and different task. We were pleased to see that with several careful adaptations, our DefScriber system was able to achieve strong results, with the 2nd best ranking out of 32 system peers in a combined overall measure.

As always, there is room for improvement and future

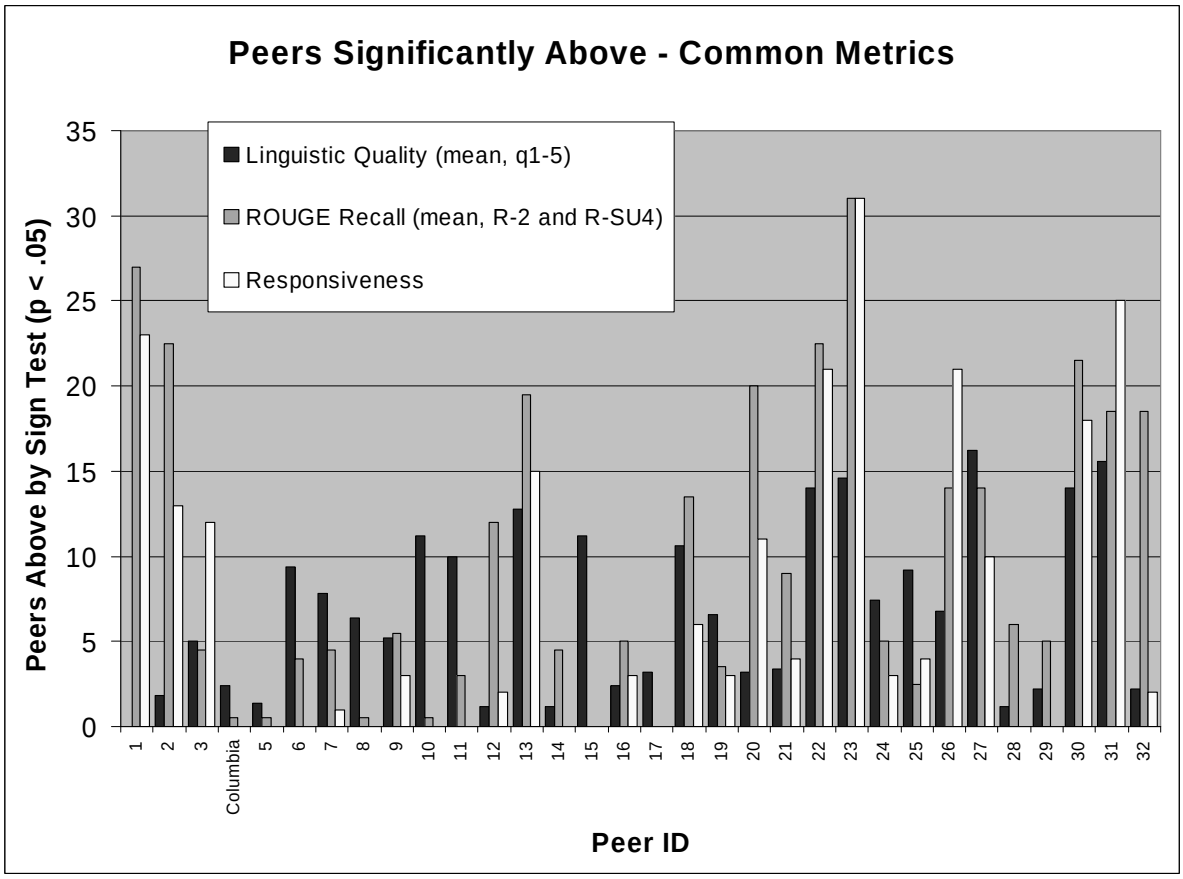

Figure 4: Systems-above rank for ROUGE, responsiveness and LQ questions.

work. In our own system, we would like to focus on improving our performance in the areas covered by the linguistic quality questions, which we believe would enhance summaries and answers for many tasks. To that end, we are currently experimenting with use of corpus-level statistics to induce rhetorical structure-like relations for improved structure/cohesion.

At the task and community level, we are enthusiastic about the continuing improvement in our understanding of various metrics which were distributed by NIST. To that end, we believe that our analysis in Section 4 demonstrates several important findings and ideas.

First, we observed a strong correlation across system rank on the three ROUGE metrics we analyzed, and also across system rank on the linguistic quality (LQ) questions. For ROUGE, this suggests taking any one of the ROUGE scores considered, or a mean, is sufficient. For the LQ questions, there was more divergence, but enough correlation to support combining the scores to get a sense of overall linguistic quality. Lastly, we introduced as a combined measure a simple mean of rank scores derived from LQ, responsiveness and ROUGE. (Further, we believe that pyramid scores might be added into this measure if more data were to be available, such that all systems were covered and differences more significant.) While this overall measure is not meant to be the last word, we feel it is an important step forward in helping give a high-level picture of system performance across the quantity and variety of metrics assessed.

6 Acknowledgments

We would like to acknowledge the support of this work by the DARPA TIDES program and ARDA AQUAINT program (contract MDA908-02-C-0008). In addition, we are thankful for the support of our colleagues in the Natural Language Processing group at Columbia University, particularly those who helped in preparing our participation in DUC 2005: David Evans, Elena Filatova, Kathleen McKeown, Ani Nenkova, Barry Schiffman and Advaith Siddarthan.

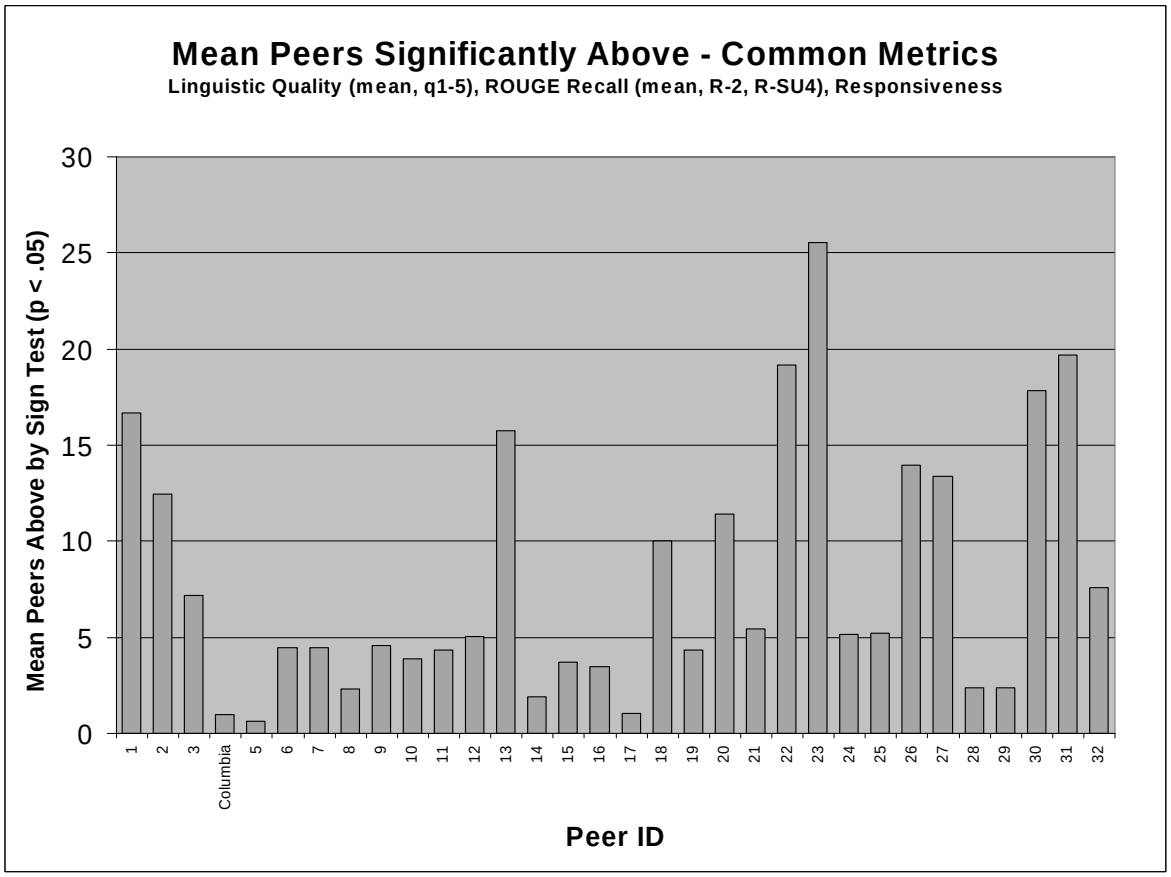

Figure 5: Mean systems-above rank across responsiveness, mean of ROUGE-2 and ROUGE-SU4 recall, and mean of linguistic quality questions 1-5.

References

[Blair-Goldensohn et al., 2004a] Blair-Goldensohn, S., Evans, D., Hatzivassiloglou, V., McKeown, K., Nenkova, A., Passonneau, R., Schiffman, B., Schlaikjer, A., Siddharthan, A., and Siegelman, S. (2004a). Columbia University at DUC 2004. In *4th Document Understanding Conference (DUC 2004) at HLT/NAACL 2004, Boston, MA*.

[Blair-Goldensohn et al., 2004b] Blair-Goldensohn, S., McKeown, K., and Schlaikjer, A. (2004b). Answering definitional questions: A hybrid approach. In Maybury, M., editor, *New Directions In Question Answering*, chapter 4. AAAI Press.

[Hovy and Lin, 1997] Hovy, E. and Lin, C. (1997). Automated text summarization in SUMMARIST. pages 18–24. In ACL '97 workshop on Intelligent Scalable Text Summarization.

[Nenkova, 2005] Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. In *20th National Conference on Artificial Intelligence (AAAI 2005)*.

[Nenkova and McKeown, 2003] Nenkova, A. and McKeown, K. (2003). References to named entities: A corpus study. In *NAACL-HLT 2003*. short paper.

[Radev et al., 2000] Radev, D. R., Jing, H., and Budzikowska, M. (2000). Centroid-based summarization of multiple documents. In ANLP-NAACL workshop on summarization.