Topic-Focused Multi-document Summarization Using an Approximate Oracle Score

John M. Conroy, Judith D. Schlesinger
IDA Center for Computing Sciences
Bowie, Maryland, USA
conroy@super.org, judith@super.org

Dianne P. O'Leary
University of Maryland
College Park, Maryland, USA
oleary@cs.umd.edu

Abstract

We consider the problem of producing a multi-document summary given a collection of documents. Since most successful methods of multi-document summarization are still largely extractive, in this paper, we explore just how well an extractive method can perform. We introduce an "oracle" score, based on the probability distribution of unigrams in human summaries. We then demonstrate that with the oracle score, we can generate extracts which score, on average, better than the human summaries, when evaluated with ROUGE. In addition, we introduce an approximation to the oracle score which produces a system with the best known performance for the 2005 Document Understanding Conference (DUC) evaluation.

1 Introduction

We consider the problem of producing a multi-document summary given a collection of documents. Most automatic methods of multi-document summarization are largely extractive. This mimics the behavior of humans for single document summarization; (Kupiec, Pendersen, and Chen 1995) reported that 79% of the sentences in a human-generated abstract were a "direct match" to a sentence in a document. In contrast, for multi-document summarization, (Copeck and Szpakowicz 2004) report that no more than 55% of the vocabulary contained in human-generated abstracts can be found in the given documents. Furthermore, multiple human summaries on the same collection of documents often have little agreement. For example, (Hovy and Lin 2002) report that unigram overlap is around 40%. (Teufel and van Halteren 2004) used a "factoid" agreement analysis of human summaries for a single document and concluded that a resulting consensus summary is stable only if 30–40 summaries are collected.

In light of the strong evidence that nearly half of the terms in human-generated multi-document abstracts are not from the original documents, and that agreement of vocabulary among human abstracts is only about 40%, we pose two coupled questions about the quality of summaries that can be attained by document extraction:

1. Given the sets of unigrams used by four human summarizers, can we produce an extract summary that is statistically indistinguishable from the human abstracts when measured by current automatic evaluation methods such as ROUGE?

2. If such unigram information can produce good summaries, can we replace this information by a statistical model and still produce good summaries?

We will show that the answer to the first question is, indeed, yes and, in fact, the unigram set information gives rise to extract summaries that usually score better than the 4 human abstractors! Secondly, we give a method to statistically approximate the set of unigrams and find it produces extracts of the DUC 05 data which outperform all known evaluated machine entries. We conclude with experiments on the extent that redundancy removal improves extracts, as well as a method of moving beyond simple extracting by employing shallow parsing techniques to shorten the sentences prior to selection.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 152–159,
Sydney, July 2006.

2 The Data

The 2005 Document Understanding Conference (DUC 2005) data used in our experiments is partitioned into 50 topic sets, each containing 25–50 documents. A topic for each set was intended to mimic a real-world complex questioning-answering task for which the answer could not be given in a short "nugget." For each topic, four human summarizers were asked to provide a 250-word summary of the topic. Topics were labeled as either "general" or "specific". We present an example of one of each category.

`Set d408c`
Granularity: Specific
Title: Human Toll of Tropical Storms
Narrative: What has been the human toll in death or injury of tropical storms in recent years? Where and when have each of the storms caused human casualties? What are the approximate total number of casualties attributed to each of the storms?

`Set d436j`
Granularity: General
Title: Reasons for Train Wrecks
Narrative: What causes train wrecks and what can be done to prevent them? Train wrecks are those events that result in actual damage to the trains themselves not just accidents where people are killed or injured.

For each topic, the goal is to produce a 250-word summary. The basic unit we extract from a document is a sentence.

To prepare the data for processing, we segment each document into sentences using a POS (part-of-speech) tagger, NLProcessor (http://www.infogistics.com/posdemo.htm). The newswire documents in the DUC 05 data have markers indicating the regions of the document, including titles, bylines, and text portions. All of the extracted sentences in this study are taken from the text portions of the documents only.

We define a "term" to be any "non-stop word." Our stop list contains the 400 most frequently occurring English words.

3 The Oracle Score

Recently, a crisp analysis of the frequency of content words used by humans relative to the high frequency content words that occur in the relevant documents has yielded a simple and powerful summarization method called SumBasic (Nenkova and Vanderwende, 2005). SumBasic produced extract summaries which performed nearly as well as the best machine systems for generic 100 word summaries, as evaluated in DUC 2003 and 2004, as well as the Multi-lingual Summarization Evaluation (MSE 2005).

Instead of using term frequencies of the corpus to infer highly likely terms in human summaries, we propose to directly model the *set* of terms (vocabulary) that is likely to occur in a sample of human summaries. We seek to estimate the probability that a term will be used by a human summarizer to first get an estimate of the best possible extract and later to produce a statistical model for an extractive summary system. While the primary focus of this work is "task oriented" summaries, we will also address a comparison with SumBasic and other systems on generic multi-document summaries for the DUC 2004 dataset in Section 8.

Our extractive summarization system is given a topic, τ, specified by a text description. It then evaluates each sentence in each document in the set to determine its appropriateness to be included in the summary for the topic τ.

We seek a statistic which can score an individual sentence to determine if it should be included as a candidate. We desire that this statistic take into account the great variability that occurs in the space of human summaries on a given topic τ. One possibility is to simply judge a sentence based upon the expected fraction of the "human summary"-terms that it contains. We posit an oracle, which answers the question "Does human summary i contain the term t?"

By invoking this oracle over the set of terms and a sample of human summaries, we can readily compute the expected fraction of human summary-terms the sentence contains. To model the variation in human summaries, we use the oracle to build a probabilistic model of the space of human abstracts. Our "oracle score" will then compute the expected number of summary terms a sentence contains, where the expectation is taken from the space of all human summaries on the topic τ.

We model human variation in summary generation with a unigram bag-of-words model on the terms. In particular, consider $P(t|\tau)$ to be the probability that a human will select term t in a summary given a topic τ. The oracle score for a sentence x, $\omega(x)$, can then be defined in terms of

P :

$$\omega(x) = \frac{1}{|x|} \sum_{t \in T} x(t) P(t|\tau)$$

where $|x|$ is the number of distinct terms sentence x contains, T is the universal set of all terms used in the topic τ and $x(t) = 1$ if the sentence x contains the term t and 0 otherwise. (We affectionally refer to this score as the "Average Jo" score, as it is derived the average uni-gram distribution of terms in human summaries.)

While we will consider several approximations to $P(t|\tau)$ (and, correspondingly, ω), we first explore the maximum-likelihood estimate of $P(t|\tau)$ given by a sample of human summaries. Suppose we are given h sample summaries generated independently. Let $c_{it}(\tau) = 1$ if the i-th summary contains the term t and 0 otherwise. Then the maximum-likelihood estimate of $P(t\tau)$ is given by

$$\hat{P}(t|\tau) = \frac{1}{h} \sum_{i=1}^{h} c_{it}(\tau).$$

We define $\hat{\omega}$ by replacing P with $\hat{P}$ in the definition of ω. Thus, $\hat{\omega}$ is the maximum-likelihood estimate for ω, given a set of h human summaries.

Given the score $\hat{\omega}$, we can compute an extract summary of a desired length by choosing the top scoring sentences from the collection of documents until the desired length (250 words) is obtained. We limit our selection to sentences which have 8 or more distinct terms to avoid selecting incomplete sentences which may have been tagged by the sentence splitter.

Before turning to how well our idealized score, $\hat{\omega}$, performs on extract summaries, we first define the scoring mechanism used to evaluate these summaries.

4 ROUGE

The state-of-the-art automatic summarization evaluation method is ROUGE (Recall Oriented Understudy for Gisting Evaluation, (Hovy and Lin 2002)), an n-gram based comparison that was motivated by the machine translation evaluation metric, Bleu (Papineni et. al. 2001). This system uses a variety of n-gram matching approaches, some of which allow gaps within the matches as well as more sophistcated analyses. Surprisingly, simple unigram and bigram matching works extremely well. For example, at DUC 05, ROUGE-2 (bigram match) had a Spearman correlation of 0.95 and a Pearson correlation of 0.97 when compared with human evaluation of the summaries for responsiveness (Dang 2005). ROUGE-n for matching $n-$grams of a summary X against h model human summaries is given by:

$$R_n(X) = \frac{\sum_{j=1}^{h} \sum_{i \in N_n} \min(X_n(i), M_n(i,j))}{\sum_{j=1}^{h} \sum_{i \in N_n} M_n(i,j),}$$

where $X_n(i)$ is the count of the number of times the n-gram i occurred in the summary and $M_n(i,j)$ is the number of times the n-gram i occurred in the j-th model (human) summary. (Note that for brevity of notation, we assume that lemmatization (stemming) is done *apriori* on the terms.)

When computing ROUGE scores, a jackknife procedure is done to make comparison of machine systems and humans more amenable. In particular, if there are k human summaries available for a topic, then the ROUGE score is computed for a human summary by comparing it to the remaining $k - 1$ summaries, while the ROUGE score for a machine summary is computed against all k subsets of size $k - 1$ of the human summaries and taking the average of these k scores.

5 The Oracle or *Average Jo* Summary

We now present results on the performance of the oracle method as compared with human summaries. We give the ROUGE-2 (R_2) scores as well as the 95% confidence error bars. In Figure 1, the human summarizers are represented by the letters A–H, and systems 15, 17, 8, and 4 are the top performing machine summaries from DUC 05. The letter "O" represents the ROUGE-2 scores for extract summaries produced by the oracle score, $\hat{\omega}$. Perhaps surprisingly, the oracle produced extracts which performed better than the human summaries! Since each human only summarized 10 document clusters, the human error bars are larger. However, even with the large error bars, we observe that the mean ROUGE-2 scores for the oracle extracts exceeds the 95% confidence error bars for several humans.

While the oracle was, of course, given the unigram term probabilities, its performance is notable on two counts. First, the evaluation metric scored on 2-grams, while the oracle was only given unigram information. In a sense, optimizing for ROUGE-1 is a "sufficient statistic" scoring at the human level for ROUGE-2. Second, the humans wrote abstracts while the oracle simply did extracting. Consequently, the documents contain sufficient text to produce human-quality extract summaries as measured by ROUGE. The human performance ROUGE scores indicate that this approach is capable of producing automatic extractive summaries that produce vocabulary comparable to that chosen by humans. Human evaluation (which we have not yet performed) is required to determine to what extent this high ROUGE-2 performance is indicative of high quality summaries for human use.

The encouraging results of the oracle score naturally lead to approximations, which, perhaps, will give rise to strong machine system performance. Our goal is to approximate $P(t|\tau)$, the probability that a term will be used in a human abstract. In the next section, we present two approaches which will be used in tandem to make this approximation.

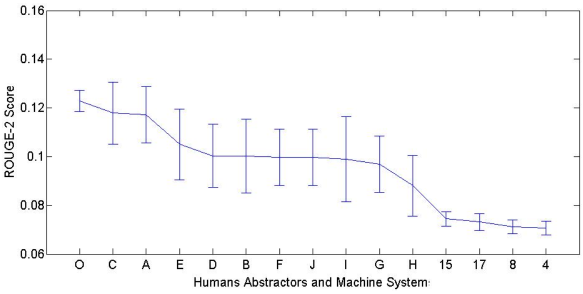

Figure 1: The Oracle (Average Jo score) Score $\hat{\omega}$

6 Approximating $P(t|\tau)$

We seek to approximate $P(t|\tau)$ in an analogous fashion to the maximum-likelihood estimate $\hat{P}(t|\tau)$. To this end, we devise methods to isolate a subset of terms which would likely be included in the human summary. These terms are gleaned from two sources, the topic description and the collection of documents which were judged relevant to the topic. The former will give rise to *query terms* and the latter to *signature terms*.

6.1 Query Term Identification

A set of query terms is automatically extracted from the given topic description. We identified individual words and phrases from both the <topic> (Title) tagged paragraph as well as whichever of the <narr> (Narrative) tagged paragraphs occurred in the topic description. We made no use of the <granularity> paragraph marking. We tagged the topic description using the POS-tagger, NLProcessor (http://www.infogistics.com/posdemo.htm), and any words that were tagged with any NN (noun), VB (verb), JJ (adjective), or RB (adverb) tag were included in a list of words to use as query terms. Table 1 shows a list of query terms for our two illustrative topics.

Set d408c: approximate, casualties, death, human, injury, number, recent, storms, toll, total, tropical, years
Set d436j: accidents, actual, causes, damage, events, injured, killed, prevent, result, train, train wrecks, trains, wrecks

Table 1: Query Terms for “Tropical Storms” and “Train Wrecks” Topics

The number of query terms extracted in this way ranged from a low of 3 terms for document set d360f to 20 terms for document set d324e.

6.2 Signature Terms

The second collection of terms we use to estimate $P(t|\tau)$ are signature terms. Signature terms are the terms that are more likely to occur in the document set than in the background corpus. They are generally indicative of the content contained in the collection of documents. To identify these terms, we use the log-likelihood statistic suggested by Dunning (Dunning 1993) and first used in summarization by Lin and Hovy (Hovy and Lin 2000). The statistic is equivalent to a mutual information statistic and is based on a 2-by-2 contingency table of counts for each term. Table 2 shows a list of signature terms for our two illustrative topics.

6.3 An estimate of $P(t|\tau)$

To estimate $P(t|\tau)$, we view both the query terms and the signature terms as “samples” from idealized human summaries. They represent the terms that we would most likely see in a human summary. As such, we expect that these sample terms may approximate the underlying set of human summary terms. Given a collection of query terms and signature terms, we can readily estimate our target objective, $P(t|\tau)$ by the following:

$$P_{qs}(t|\tau) = \frac{1}{2}q_t(\tau) + \frac{1}{2}s_t(\tau)$$

`Set d408c:` ahmed, allison, andrew, bahamas, bangladesh, bn, caribbean, carolina, caused, cent, coast, coastal, croix, cyclone, damage, destroyed, devastated, disaster, dollars, drowned, flood, flooded, flooding, floods, florida, gulf, ham, hit, homeless, homes, hugo, hurricane, insurance, insurers, island, islands, lloyd, losses, louisiana, manila, miles, nicaragua, north, port, pounds, rain, rains, rebuild, rebuilding, relief, remnants, residents, roared, salt, st, storm, storms, supplies, tourists, trees, tropical, typhoon, virgin, volunteers, weather, west, winds, yesterday.
`Set d436j:` accident, accidents, ammunition, beach, bernardino, board, boulevard, brake, brakes, braking, cab, car, cargo, cars, caused, collided, collision, conductor, coroner, crash, crew, crossing, curve, derail, derailed, driver, emergency, engineer, engineers, equipment, fe, fire, freight, grade, hit, holland, injured, injuries, investigators, killed, line, locomotives, maintenance, mechanical, miles, morning, nearby, ntsb, occurred, officials, pacific, passenger, passengers, path, rail, railroad, railroads, railway, routes, runaway, safety, san, santa, shells, sheriff, signals, southern, speed, station, train, trains, transportation, truck, weight, wreck

Table 2: Signature Terms for "Tropical Storms" and "Train Wrecks" Topics

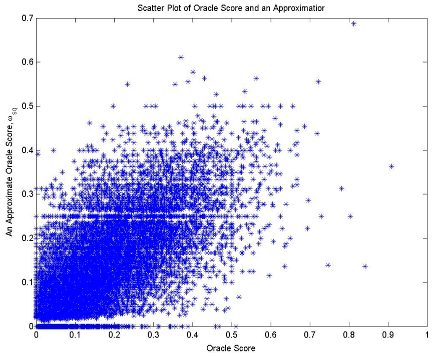

Figure 2: Scatter Plot of $\hat{\omega}$ versus ω_{qs}

where $s_t(\tau)$=1 if t is a signature term for topic τ and 0 otherwise and $q_t(\tau) = 1$ if t is a query term for topic τ and 0 otherwise.

More sophisticated weightings of the query and signature have been considered; however, for this paper we limit our attention to the above elementary scheme. (Note, in particular, a psuedo-relevance feedback method was employed by (Conroy et. al. 2005), which gives improved performance.)

Similarly, we estimate the oracle score of a sentence's expected number of human abstract terms as

$$\omega_{qs}(x) = \frac{1}{|x|} \sum_{t \in T} x(t) P_{qs}(t|\tau)$$

where $|x|$ is the number of distinct terms that sentence x contains, T is the universal set of all terms and $x(t) = 1$ if the sentence x contains the term t and 0 otherwise.

For both the oracle score and the approximation, we form the summary by taking the top scoring sentences among those sentences with at least 8 distinct terms, until the desired length (250 words for the DUC05 data) is achieved or exceeded. (The threshold of 8 was based upon previous analysis of the sentence splitter, which indicated that sentences shorter than 8 terms tended not be be well formed sentences or had minimal, if any, content.) If the length is too long, the last sentence chosen is truncated to reach the target length.

Figure 2 gives a scatter plot of the oracle score ω and its approximation ω_{qs} for all sentences with at least 8 unique terms. The overall Pearson correlation coefficient is approximately 0.70. The correlation varies substantially over the topics. Figure 3 gives a histogram of the Pearson correlation coefficients for the 50 topic sets.

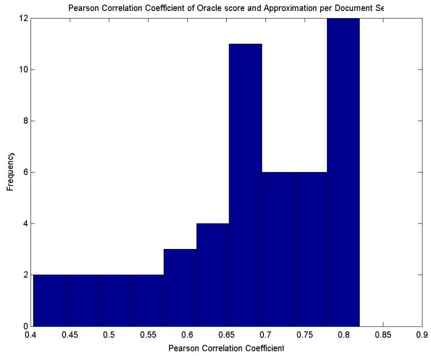

Figure 3: Histogram of Document Set Pearson Coefficients of $\hat{\omega}$ versus ω_{qs}

7 Enhancements

In the this section we explore two approaches to improve the quality of the summary, linguistic preprocessing (sentence trimming) and a redundancy removal method.

7.1 Linguistic Preprocessing

We developed patterns using "shallow parsing" techniques, keying off of lexical cues in the sentences after processing them with the POS-tagger. We initially used some full sentence eliminations along with the phrase eliminations itemized below; analysis of DUC 03 results, however, demonstrated that the full sentence eliminations were not useful.

The following phrase eliminations were made, when appropriate:

- gerund clauses;
- restricted relative-clause appositives;
- intra-sentential attribution;
- lead adverbs.

See (Dunlavy et. al) for the specific rules used for these eliminations. Comparison of two runs in DUC 04 convinced us of the benefit of applying these phrase eliminations on the full documents, prior to summarization, rather than on the selected sentences after scoring and sentence selection had been performed. See (Conroy et. al. 2004) for details on this comparison.

After the trimmed text has been generated, we then compute the signature terms of the document sets and recompute the approximate oracle scores. Note that since the sentences have usually had some extraneous information removed, we expect some improvement in the quality of the signature terms and the resulting scores. Indeed, the median ROUGE-2 score increases from 0.078 to 0.080.

7.2 Redundancy Removal

The greedy sentence selection process we described in Section 6 gives no penalty for sentences which are redundant to information already contained in the partially formed summary. A method for reducing redundancy can be employed. One popular method for reducing redundancy is maximum marginal relevance (MMR) (2). Based on previous studies, we have found that a pivoted QR, a method from numerical linear algebra, has some advantages over MMR and performs somewhat better.

Pivoted QR works on a term-sentence matrix formed from a set of candidate sentences for inclusion in the summary. We start with enough sentences so the total number of terms is approximately twice the desired summary length. Let B be the term-sentence matrix with $B_{ij} = 1$ if sentence j contains term i.

The columns of B are then normalized so their 2-norm (Euclidean norm) is the corresponding approximate oracle score, i.e. $\omega_{qs}(b_j)$, where b_j is the j-th column of B. We call this normalized term sentence matrix A.

Given a normalized term-sentence matrix A, QR factorization attempts to select columns of A in the order of their importance in spanning the subspace spanned by all of the columns. The standard implementation of pivoted QR decomposition is a "Gram-Schmidt" process. The first r sentences (columns) selected by the pivoted QR are used to form the summary. The number r is chosen so that the summary length is close to the target length. A more complete description can be found in (Conroy and O'Leary 2001).

Note, that the selection process of using the pivoted QR on the weighted term sentence matrix will first choose the sentence with the highest ωpq score as was the case with the greedy selection process. Its subsequent choices are affected by previous choices as the weights of the columns are decreased for any sentence which can be approximated by a linear combination of the current set of selected sentences. This is more general than simply demanding that the sentence have small overlap with the set of previous chosen sentences as

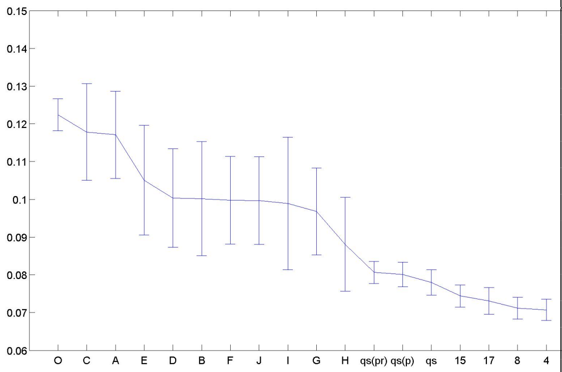

Figure 4: ROUGE-2 Performance of Oracle Score Approximations $\hat{\omega}$ vs. Humans and Peers

Submission	Mean	95% CI Lower	95% CI Upper
O (ω)	0.13710	0.13124	0.14299
C	0.13260	0.11596	0.15197
D	0.12380	0.10751	0.14003
B	0.11788	0.10501	0.13351
G	0.11324	0.10195	0.12366
F	0.10893	0.09310	0.12780
H	0.10777	0.09833	0.11746
J	0.10717	0.09293	0.12460
I	0.10634	0.09632	0.11628
E	0.10365	0.08935	0.11926
A	0.10361	0.09260	0.11617
24	0.09558	0.09144	0.09977
$\omega_{qs}^{(pr)}$	0.09160	0.08729	0.09570
15	0.09097	0.08671	0.09478
12	0.08987	0.08583	0.09385
8	0.08954	0.08540	0.09338
23	0.08792	0.08371	0.09204
$\omega_{qs}^{(p)}$	0.08738	0.08335	0.09145
ω_{qs}	0.08713	0.08317	0.09110
28	0.08700	0.08332	0.09096

Table 3: Average ROUGE 2 Scores for DUC06: Humans A-I

would be done using MMR.

8 Results

Figure 4 gives the ROUGE-2 scores with error bars for the approximations of the oracle score as well as the ROUGE-2 scores for the human summarizers and the top performing systems at DUC 2005. In the graph, qs is the approximate oracle, qs(p) is the approximation using linguistic preprocessing, and qs(pr) is the approximation with both linguistic preprocessing and redundancy removal. Note that while there is some improvement using the linguistic preprocessing, the improvement using our redundancy removal technique is quite minor. Regardless, our system using signature terms and query terms as estimates for the oracle score performs comparably to the top scoring system at DUC 05.

Table 3 gives the ROUGE-2 scores for the recent DUC 06 evaluation which was essentially the same task as for DUC 2005. The manner in which the linguistic preprocessing is performed has changed from DUC 2005, although the types of removals have remained the same. In addition, pseudo-relevance feedback was employed for redundancy removal as mentioned earlier. See (Conroy et. al. 2005) for details.

While the main focus of this study is task-oriented multidocument summarization, it is instructive to see how well such an approach would perform for a generic summarization task as with the 2004 DUC Task 2 dataset. Note, the ω score for generic summaries uses only the signature term portion of the score, as no topic description is given. We present ROUGE-1 (rather than ROUGE-2) scores with stop words removed for comparison with the published results given in (Nenkova and Vanderwende, 2005).

Table 4 gives these scores for the top performing systems at DUC04 as well as SumBasic and $\omega_{qs}^{(pr)}$, the approximate oracle based on signature terms alone with linguistic preprocess trimming and pivot QR for redundancy removal. As displayed, $\omega_{qs}^{(pr)}$ scored second highest and within the 95% confidence intervals of the top system, peer 65, as well as SumBasic, and peer 34.

Submission	Mean	95% CI Lower	95% CI Upper
F	0.36787	0.34442	0.39467
B	0.36126	0.33387	0.38754
O (ω)	0.35810	0.34263	0.37330
H	0.33871	0.31540	0.36423
A	0.33289	0.30591	0.35759
D	0.33212	0.30805	0.35628
E	0.33277	0.30959	0.35687
C	0.30237	0.27863	0.32496
G	0.30909	0.28847	0.32987
$\omega_{qs}^{(pr)}$	0.308	0.294	0.322
peer 65	0.308	0.293	0.323
SumBasic	0.302	0.285	0.319
peer 34	0.290	0.273	0.307
peer 124	0.286	0.268	0.303
peer 102	0.285	0.267	0.302

Table 4: Average ROUGE 1 Scores with stop words removed for DUC04, Task 2

9 Conclusions

We introduced an oracle score based upon the simple model of the probability that a human will choose to include a term in a summary. The oracle score demonstrated that for task-based summarization, extract summaries score as well as human-generated abstracts using ROUGE. We then demonstrated that an approximation of the oracle score based upon query terms and signature terms gives rise to an automatic method of summarization, which outperforms the systems entered in DUC05. The approximation also performed very well in DUC 06. Further enhancements based upon linguistic trimming and redundancy removal via a pivoted QR algorithm give significantly better results.

References

Jamie Carbonnell and Jade Goldstein "The of MMR, diversity-based reranking for reordering documents and producing summaries." In Proc. ACM SIGIR, pages 335–336.

John M. Conroy and Dianne P. O'Leary. "Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition". Technical report, University of Maryland, College Park, Maryland, March, 2001.

John M. Conroy and Judith D. Schlesinger and Jade Goldstein and Dianne P. O'Leary, Left-Brain Right-Brain Multi-Document Summarization, *Document Understanding Conference 2004* http://duc.nist.gov/ 2004

John M. Conroy and Judith D. Schlesinger and Jade Goldstein, CLASSY Tasked Based Summarization: Back to Basics, *Document Understanding Conference 2005* http://duc.nist.gov/ 2005

John M. Conroy and Judith D. Schlesinger Dianne P. O'Leary, and Jade Goldstein, Back to Basciss: CLASSY 2006, *Document Understanding Conference 2006* http://duc.nist.gov/ 2006

Terry Copeck and Stan Szpakowicz 2004 Vocabulary Agreement Among Model Summaries and Source Documents In *ACL Text Summarization Workshop, ACL 2004.*

Hoa Trang Dang Overview of DUC 2005 *Document Understanding Conference 2005* http://duc.nist.gov

Daniel M. Dunlavy and John M. Conroy and Judith D. Schlesinger and Sarah A. Goodman and Mary Ellen Okurowski and Dianne P. O'Leary and Hans van Halteren, "Performance of a Three-Stage System for Multi-Document Summarization", DUC 03 Conference Proceedings, http://duc.nist.gov/, 2003

Ted Dunning, "Accurate Methods for Statistics of Surprise and Coincidence", *Computational Linguistics*, 19:61-74, 1993.

Julian Kupiec,, Jan Pedersen, and Francine Chen. "A Trainable Document Summarizer". *Proceedings of the 18th Annual International SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, 1995.

Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics*, pages 495–501, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. Manual and Automatic Evaluation of Summaries In *Document Understanding Conference 2002* http:/duc.nist.gov

Multi-Lingual Summarization Evaluation http://www.isi.edu/ cyl/MTSE2005/MLSummEval.html

NLProcessor http://www.infogistics.com/posdemo.htm

Ani Nenkova and Lucy Vanderwende. 2005. *The Impact of Frequency on Summarization*,MSR-TR-2005-101. Microsoft Research Technical Report.

Kishore Papineni and Salim Roukos and Todd Ward and Wei-Jing Zhu, Bleu: a method for automatic evaluation of machine translation, *Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center (2001)*

Simone Teufel and Hans van Halteren. 2004. *4: Evaluating Information Content by Factoid Analysis: Human Annotation and Stability*, EMNLP-04, Barcelona