Automatic Summary Evaluation without Human Models

Annie Louis
University of Pennsylvania
Philadelphia, PA 19104, USA
lannie@seas.upenn.edu

Ani Nenkova
University of Pennsylvania
Philadelphia, PA 19104, USA
nenkova@seas.upenn.edu

Abstract

We present a fully automatic approach for summarization evaluation that does not require the creation of human model summaries.[1] Our work capitalizes on the fact that a summary contains the most representative information from the input and so it is reasonable to expect that the distribution of terms in the input and a good summary are similar to each other. To compare the term distributions, we use KL and Jensen-Shannon divergence, cosine similarity, as well as unigram and multinomial models of text. Our results on a large scale evaluation from the Text Analysis Conference show that input-summary comparisons can be very effective. They can be used to rank participating systems very similarly to manual model-based evaluations (pyramid evaluation) as well as to manual human judgments of summary quality without reference to a model. Our best feature, Jensen-Shannon divergence, leads to a correlation as high as 0.9 with manual evaluations.

1 Introduction

Automatic text summarizers are expected to produce a condensed form of an input, retaining the most important content and presenting the information in a coherent fashion. The development of suitable and efficient evaluations of summary content and organization is crucial and necessary to guide system development. Ideally, summary quality would be measured through extrinsic evaluations where the usefulness of a system summary is assessed in specific task scenarios such as reading comprehension (Morris et al., 1992), relevance judgements in information retrieval (Mani et al., 2002; Jing et al., 1998; Brandow et al., 1995) and other tasks (McKeown et al., 2005; Sakai and Sparck-Jones, 2001; Mani et al., 2002; Tombros and Sanderson, 1998; Roussinov and Chen, 2001). However, organizing and carrying out such evaluations is difficult and in practice intrinsic evaluations are the standard in summarization. In intrinsic evaluations, system summaries are either compared with reference summaries produced by humans (model summaries), or directly assessed by human judges on a scale (most commonly 1 to 5), without reference to a model summary. The refinement and usability analysis of these evaluation techniques have been the focus of large scale evaluation efforts such as the Document Understanding Conferences (DUC) (Baldwin et al., 2000; Harman and Over, 2004; Over et al., 2007) and TIPSTER SUMMAC reports (Mani et al., 2002).

Still, in recent years by far the most popular evaluation method used during system development and for reporting results in publications has been the automatic evaluation tool ROUGE (Lin, 2004; Lin and Hovy, 2003). ROUGE compares system summaries against one or more model summaries automatically, by computing n-gram word overlaps between the two. The wide adoption of such automatic measures is understandable, as they are convenient and have greatly reduced the complexity of evaluations. They have also been shown to correlate well with manual evaluations of content, based on comparison with a single model summary, as used in the early editions of DUC.

However, the creation of gold standard summaries for comparison is still time-consuming and expensive. In our work, we explore the feasibility of developing *a fully automatic* evaluation method, that does not make use of human model summaries at all. Proposals for developing such fully automatic methods have been put forward in the past, but no substantial progress has been made so far in this research direction.

For example in (Radev et al., 2003), a large scale fully automatic evaluation of eight summarizer systems on 18,000 documents was performed without any human

[1] This work was presented at the poster session in TAC on Nov. 18, 2008. We would like to thank Hoa Trang Dang and the TAC advisory board for giving us the opportunity to work on this project.

effort, using an information retrieval scenario. A search engine was used to rank documents according to their relevance to a posed query. The summaries for each document were also ranked for relevance with respect to the same query. For good summarization systems, the ranking of relevance of summaries is expected to be similar to that of the full documents. Based on this intuition, the correlation between query relevance rankings of a system's summaries and the ranking of original documents was used to compare the different systems. Effectively this approach is motivated by the assumption that the distribution of terms in a good summary is similar to the distribution of terms in the original input text.

Even earlier, (Donaway et al., 2000) suggested that there are considerable benefits to be had in adopting model-free methods of evaluation involving direct comparisons between input and summary. Their work was motivated by the well documented fact that there are multiple good summaries of the same text and that there is incredible variation in content selection choices in human summarization (Rath et al., 1961; Radev and Tam, 2003; van Halteren and Teufel, 2003; Nenkova and Passonneau, 2004). As a result, the identity of the model writer significantly affects summary evaluations (McKeown et al., 2001), and evaluations of the same systems can be rather different when two different models are used. In their experiments, Donaway et al. demonstrated that the correlation between a manual evaluation using comparison with a model summary and a) manual evaluation using comparison with a different model summary and b) evaluation by directly comparing input and summary, are the same. They used cosine similarity with singular value decomposition to perform the input-summary comparison. Their conclusion was that automatic methods for comparison between input and a summary should be seriously considered as an alternative for evaluation.

In this paper, we present a comprehensive study of fully automatic summary evaluation, without human models. A summary's content is judged for quality by directly estimating its closeness to the input. We compare several probabilistic (information-theoretic) approaches for characterizing the similarity and differences between input and summary content. The utility of the various approaches for comparison varies widely, but a number of them lead to rankings of systems that correlate well with manual evaluations performed in the recent Text Analysis Conference (NIST). A simple information theoretic measure, Jensen Shannon divergence between input and summary emerges as the best feature. System rankings produced using this measure lead to correlations with human rankings as high as 0.9.

2 Data: TAC 2008

2.1 Topic focused and Update Summarization

Two types of summaries, query focused summaries and update summaries, were evaluated in the main task of the summarization track of the 2008 Text Analysis Conference (TAC) run by NIST[2]. Query focused summaries are produced from the input documents in response to a stated user information need (query). The update summaries require more sophistication: two sets of articles on the same topic are provided. The first set of articles represent the background of a story and the users are assumed to be already familiar with the information contained in them. The task is produce a multi-document summary from the second set of articles on the same topic, that can serve as an update to the user. This task is reminiscent of the novelty detection task explored at TREC (Soboroff and Harman, 2005).

2.2 Test set

The test set for the TAC 2008 update task contains 48 inputs. Each input consists of two sets of documents, A and B, of 10 documents each. A and B are on the same general topic, and B contains documents published later than those in A. In addition, for each input, the user's need is represented by a topic statement which consists of a title and narrative. An example topic statement is given below.

Title: Airbus A380
Narrative: Describe developments in the production and launch of the Airbus A380.

The task for participating systems is to produce two summaries, a query focused summary of document set A and an update summary of document set B. The first summary is expected to summarize the documents in A using the topic statement to focus content selection. The second summary is expected to be a compilation of updates from document set B, assuming that the user has read all the documents in A. The maximum word limit for both types of summaries is 100 words.

In order to allow for in-depth discussion, we will analyze our findings only for query focused summaries. Similar results were obtained for the evaluation of update summaries and are reported in separate tables in Section 6.

2.3 Summarizers

There were 57 participating systems in TAC 2008. The baseline summary in both cases— query focused and update tasks was created by choosing the first sentences of

[2] http://www.nist.gov/tac

manual score	R-1 recall	R-2 recall
Query Focused summaries		
pyramid score	0.859	0.905
responsiveness	0.806	0.873
Update summaries		
pyramid score	0.912	0.941
responsiveness	0.865	0.884

Table 1: Spearman correlation between system scores assigned by manual methods and those assigned by ROUGE, ROUGE-1 and ROUGE-2 recall (TAC2008, 57 systems). All correlations are highly significant with p-value < 0.00001.

the most recent document in document sets A and B respectively. In addition, four human summaries were produced for each type to serve as model summaries for evaluation. Only the 57 systems were used for our evaluation experiments.

2.4 Evaluations

Both manual and automatic evaluations were conducted at NIST to assess quality of summaries produced by the systems. Summarizer performance is defined by two key aspects of summary quality— content selection (identification of important content in the input) and linguistic quality (structure and presentation of selected content). Three methods of manual evaluation were used to assign scores to summaries.

Pyramid eval The pyramid evaluation method (Nenkova and Passonneau, 2004) has been developed for reliable and diagnostic assessment of content selection quality in summarization and has been used in several large scale evaluations (Nenkova et al., 2007). It uses multiple human models from which annotators identify semantically defined Summary Content Units (SCU). Each SCU is assigned a weight equal to the number of human model summaries that express that SCU. An ideal maximally informative summary would express a subset of the most highly weighted SCUs, with multiple maximally informative summaries being possible.

Four human summaries were used for the annotation. The pyramid score for a system summary is equal to the ratio between the sum of weights of SCUs expressed in a summary (again identified manually) and the sum of weights of an ideally informative summary with the same number of SCUs.[3]

[3]In addition, pyramids using all combinations of three models were constructed from the four-model pyramid. A human summary was scored against a pyramid comprising of SCUs from the other three model summaries. Jackknifing was implemented for system summaries by comparing them to each of the four 3-model pyramids obtaining a pyramid score from each comparison. The average of these scores is also reported together with the score from the four-model pyramid. The correlation between the two pyramid scores (using three and four models) is very high— 0.9997 for query focused and 0.9993 for update tasks. So we will not discuss the correlations of our features with the three-model pyramid scores.

Responsiveness eval The responsiveness of a summary is defined as a measure of overall quality combining content selection and linguistic quality: summaries must present useful content in a structured fashion in order to better satisfy the user's need. Assessors directly assigned responsiveness scores on a scale of 1 (poor summary) to 5 (very good summary) to each summary. The assessments are done without reference to any model summaries.

ROUGE NIST also evaluated the summaries automatically using ROUGE (Lin, 2004; Lin and Hovy, 2003). Comparison between a summary and a set of model summaries is computed using unigram (R1) and bigram overlaps (R2). The scores were computed after stemming but stop words were retained in the summaries. Table 1 shows that ROUGE obtains very good correlations with the manual scores for content selection. The correlation with pyramid scores is 0.90 and 0.94 for query focused and update summary respectively, and 0.87 and 0.88 with responsiveness. Given these observations, ROUGE is a high performance automatic evaluation metric when human summaries are available and sets a high standard for comparison of other automatic evaluation methods.

Linguistic quality questions were used to assess readability and well-formedness of the produced summaries. Assessors scored the well-formedness of a summary on a scale from 1 (very poor) to 5 (very good). Grammaticality, non-redundancy, referential clarity, focus, structure and coherence were the factors to be considered during evaluation.

But since our features are designed to capture content selection quality, manual pyramid and responsiveness scores will be used for comparison with our automatic method. The correlation between these two evaluations is overall high, 0.885 and 0.923 respectively for query focused and update summarization. Despite of this, we use both measures as a reference for comparison with our fully automatic evaluation method because albeit high, the correlation between them is not perfect (as was the correlation between the two alternative pyramid scores).

3 Features

We describe three classes of features used to compare input and summary content: distributional similarity, summary likelihood and use of topic signatures. Words in both input and summary were stemmed before feature computations.

3.1 Distributional Similarity

Measures of similarity between two probability distributions are a natural choice for the task at hand. We choose

to experiment with three common such measures: KL and Jensen Shannon divergence and cosine similarity. We expect that good summaries are characterized by low divergence between the probability distributions of words in the input and the summary, and by high similarity with the input.

Moreover, these three metrics have already been used for summary evaluation, albeit in different contexts. (Lin et al., 2006) compared the performance of ROUGE with KL and JS divergence for the evaluation of summaries using human models. The divergence between human and machine summary distributions was used as an estimate of summary score. The study found that JS divergence always outperformed KL divergence and using multiple human references, the performance of JS divergence was better than standard ROUGE scores for multi-document summarization. JS divergence has also been found useful in other NLP tasks as a good predictor of unseen events (Dagan et al., 1994; Lapata, 2000).

The use of cosine similarity in (Donaway et al., 2000) is more directly related to our work. In this study, it was shown that the differences between evaluations based on two different models is about the same as the difference between system ranking based on one model summary and ranking produced using input-summary comparisons. Cosine similarity with singular value decomposition was used to compare input with summaries. Only this one approach for similarity comparison was used. In contrast, we explore a variety of features and the experiments outlined in this paper enable us to compare the usefulness of different similarity measures for automatic evaluation.

Kullback Leibler (KL) divergence The KL divergence between two probability distributions P and Q is given by

$$D(P||Q) = \sum_{w} p_P(w) \log_2 \frac{p_P(w)}{p_Q(w)} \quad (1)$$

It is defined as the average number of bits wasted by coding samples belonging to P using another distribution Q, an approximate of P. In our case, the two distributions are those for words in the input and summary respectively. However, KL divergence is not symmetric. So the divergence computed both ways, input-summary and summary-input are used as features.

In addition, the divergence is undefined when $p_P(w) > 0$ but $p_Q(w) = 0$. We perform simple smoothing to overcome the problem.

$$p(w) = \frac{C + \delta}{N + \delta * B} \quad (2)$$

Here C is the count of word w and N is the number of tokens. A value of 1.5 times the input vocabulary was used as an estimate for outcomes (B) of the probability distribution and δ was set to a small value of 0.0005 to avoid shifting too much probability mass to unseen events.

Jensen Shannon (JS) divergence The JS divergence is based on the idea that the distance between two distributions cannot be very different from the average of distances from their mean distribution. It is formally defined as

$$J(P||Q) = \frac{1}{2}[D(P||A) + D(Q||A)],$$

where $A = \frac{P+Q}{2}$ is the mean distribution of P and Q.

In contrast to KL divergence, the JS distance is symmetric and always defined. We use both smoothed and unsmoothed versions of the divergence as features.

Similarity between input and summary The third metric is cosine overlap between the tf-idf vector representations of input and summary contents.

$$cos\theta = \frac{v_{inp}.v_{summ}}{||v_{inp}||||v_{summ}||}. \quad (3)$$

We compute two variants,

1. Cosine overlap between input and summary words
2. Cosine overlap between topic signatures of input and words of summary

Topic signatures are words highly descriptive of the input, as determined by the application of log-likelihood test (Lin and Hovy, 2000). Using only topic signatures from the input to represent text is expected to be more accurate and to remove noise from peripherally related content. In addition, the refined input vector has a smaller dimension suitable for comparison with a vector of summary words which is typically small compared to a complete bag of words vector of the input.

3.2 Summary Probabilities

These features capture the log likelihood of a summary given its input. The probability of a word appearing in the summary is estimated from the input. We compute both the unigram bag of words probability as well as the probability of the summary under a multinomial model. The comparison with ROUGE in (Lin et al., 2006) (described under Section 3.1) also included unigram log likelihood alongside KL and JS divergences. However JS divergence proved better than the other two.

Unigram summary probability

$$(p_{inp}w_1)^{n_1}(p_{inp}w_2)^{n_2}...(p_{inp}w_r)^{n_r} \quad (4)$$

where $p_{inp}w_i$ is the probability in the input of word w_i, n_i is the number of times w_i appears in the summary, and $w_1...w_r$ are all words in the summary vocabulary.

Multinomial summary probability

$$\frac{N!}{n_1!n_2!...n_r!}(p_{inp}w_1)^{n_1}(p_{inp}w_2)^{n_2}...(p_{inp}w_r)^{n_r} \quad (5)$$

where $N = n_1 + n_2 + ... + n_r$ is the total number of words in the summary.

3.3 Use of input's topic words in summary

Summarizer systems that directly optimize for more topic signatures during content selection have fared very well in evaluations (Conroy et al., 2006). Hence the number of topic signatures from the input present in a summary might be a good indicator of summary content quality. We experiment with two features that quantify the presence of topic signatures in a summary.

1. Percentage of summary composed from input's topic signatures
2. Percentage of topic signature words from the input that also appear in the summary

While both features will obtain higher values for summaries containing many topic signature words, the first is guided simply by the presence of any topic word while the second measures the diversity of topic words used in the summary.

3.4 Feature combination using linear regression

We also evaluated the performance of a feature combining all of the above features into a single measure using linear regression. The value of the feature for each summary was obtained using leave-one-out approach: for a particular input and system-summary combination, a linear regression model using the automatic features to predict the manual evaluation scores was trained. The training set consisted only of examples which included neither the same input nor the same system. Hence during training, no examples of either the test input or system were seen.

4 Comparison to manual evaluations

In this section, we report the correlations between system ranking using our automatic features and manual evaluations. More precisely, the value of features was computed for each summary submitted for evaluations. We studied the predictive power of features in two scenarios *macro level; per system*: the average feature value across all inputs was used to rank the systems. The average manual score (pyramid or responsiveness) was also computed for each system, and the correlations between the two rankings were analyzed; *micro level; per input* the systems were ranked for each input separately, and correlations between the summary ranking for each input were computed. The two levels of analysis address different questions: Can we automatically identify system performance across all test inputs (macro level) and can we identify which summaries for a given input were good and which were bad (micro level) ?

Features	**pyramid score**	**responsiveness**
JSD div	-0.880	-0.736
JSD div smoothed	-0.874	-0.737
% of ip topic in summ	0.795	0.627
KL div summ-inp	-0.763	-0.694
cosine inp-summ	0.712	0.647
% of summ = topic wd	0.712	0.602
topic overlap	0.699	0.629
KL div inp-summ	-0.688	-0.585
mult. summ prob	0.222	0.235
unigram summ prob	-0.188	-0.101
regression	0.867	0.705

Table 2: Spearman correlation between fully automatically computed features and manually assigned system scores (avg. over all test inputs) for the query focused summarization subtask in TAC 2008. All results are highly significant with p-values < 0.000001 except unigram and multinomial summary probability, which are not significant.

In addition, we compare our results to model-based evaluations using ROUGE and analyze the effects of stemming the input and summary vocabularies.

4.1 Performance at macro level

Table 2 shows the Spearman correlation between the manual and automatic scores averaged across the 48 inputs. We find that both distributional similarity as well as the topic signature features obtain rankings very similar to those produced by humans while summary probabilities turn out unsuitable for the evaluation task.

Notably, the linear regression combination of features does not lead to better results than the single best feature: the JS divergence. It outperforms other features including the regression metric and obtains the best correlations with both types of manual scores, 0.88 with pyramid score and 0.74 with responsiveness. The correlation with pyramid score is in fact better than that obtained by ROUGE-1 recall (0.86). Similar results establishing that JS divergence is the most suitable measure for automatic evaluation were reported in a study of model-based evaluation metrics (Lin et al., 2006). In their study of generic multi-document summarization, JS divergence between system and model summaries obtained better correlations with manual rankings than ROUGE overlap scores. Our results provide further evidence that this divergence metric is indeed best suited for content comparison of two texts.

The best topic signature based feature—percentage of input's topic signatures that are present in the summary—ranks next only to JS divergence and regression. Given this result, systems optimizing for topic signatures would

score well with respect to content as was observed in previous large scale evaluations conducted by NIST. We also find that the feature simply reflecting the proportion of topic signatures in the summary performs worse as an evaluation metric. This observation leads us to the conclusion that a summary that contains many different topic signatures from the input seems to carry better content than one that contains topic signatures of fewer types.

The most simple comparison metric—cosine overlap of words—performs worse than the best divergence and topic signature features. The modified overlap score of input topic signatures and summary words also fails to obtain very high correlations. The rankings based on unigram and multinomial summary probabilities do not correlate with manual scores. Almost all systems use frequency in some form to inform content selection and this could be a reason why likelihood fails to distinguish between the system summaries.

4.2 Performance on micro level

As a more stringent assessment of the automatic evaluation, let us consider the rankings obtained on a per-input basis. These results are summarized in Table 3. The number of inputs for which correlations were significant are reported along with their minimum, maximum values and the number of inputs for which correlations above 0.5 were observed. The results are less spectacular at the level of individual inputs: JS divergence rankings obtain significant correlations with pyramid scores for 73% of the inputs. Further, only 40% of inputs obtain correlations above 0.5. The results are worse for other features and for comparison with responsiveness scores.

Overall, the micro level results suggest that the fully automatic measures we examined will not be useful for providing information about summary quality for a single input. For average over many test sets, the fully automatic evaluations measures give more reliable and useful results, highly correlated with rankings produced by manual evaluations.

4.3 Effects of stemming

So far, the analysis was based on feature values computed after stemming the input and summary words. We also computed the values of the same features without stemming and found that divergence metrics benefit greatly when stemmed vocabularies are used. The biggest improvements in correlations are for JS and KL divergences with respect to responsiveness. For JS divergence, the correlation increases from 0.571 to 0.736 and for KL divergence (summary-input), from 0.528 to 0.694. Before stemming, topic signature and bag of words overlap features are best predictive of responsiveness (correlations are 0.630 and 0.642 respectively) but do not change much after stemming (topic overlap— 0.629, bag of words—0.647). Divergences emerge as better metrics only after stemming. Stemming also proves beneficial for likelihood features. Before stemming, their correlations are directed in the wrong direction, they improve after stemming to being either positive or closer to zero. However, these probabilities remain unable to produce human-like rankings.

4.4 Difference in correlations: pyramid and responsiveness scores

Overall, we find that correlations with pyramid scores are higher than correlations with responsiveness. Clearly our features are designed to compare input-summary contents only. On the other hand, higher order ROUGE n-gram scores can be expected to capture some aspects of fluency in addition to an estimate of content quality. Since responsiveness judgements were based on both content and linguistic qualities of summaries, it is not surprising that these rankings are harder to replicate using our content based features.

5 Comparison with ROUGE

Although, JS divergence outperforms ROUGE-1 recall for correlations with pyramid scores at the average level, ROUGE-2 recall is still better. Also, ROUGE obtains the best correlations with responsiveness judgements. At the per-input micro level, ROUGE clearly gives the best human-like rankings— ROUGE-1 recall obtains significant correlations for over 95% of inputs and correlations above 0.5 for at least 50% of inputs. The ROUGE results are shown in the last two rows of Table 3.

However, when making these comparisons we need to keep in mind the fact that ROUGE evaluates system summaries using four manual models for each input. The evaluations using our features are fully automatic, with no human summaries at all.

For manual pyramid scores, the best obtained correlation with fully automatic evaluation is 0.88 (JS divergence) while the best correlation with ROUGE is 0.90 (R2). The difference is negligably small for content-based evaluations.

For manual responsiveness scores which combine aspects of linguistic quality along with content selection evaluation, the best correlations are 0.73 (JS divergence) and 0.87 (R2). For this measure, the difference between ROUGE and the fully automatic comparisons is significant, indicating that our intuition that the proposed metrics will not be suitable for linguistic quality evaluation was correct. Other metrics for linguistic quality need to be explored for this task (Lapata and Barzilay, 2005).

	pyramid						responsiveness					
features	**max**	**min**	**sig**	**sig%**	**a0.5**	**a0.5%**	**max**	**min**	**sig**	**sig%**	**a0.5**	**a0.5%**
JSD	-0.714	-0.271	35	72.9	19	39.6	-0.654	-0.262	35	72.9	17	35.4
JSD smoothed	-0.712	-0.269	35	72.9	18	37.5	-0.649	-0.279	33	68.8	17	35.4
KL summ-inp	-0.736	-0.276	35	72.9	17	35.4	-0.628	-0.261	35	72.9	13	27.1
% of input sign	0.701	0.286	31	64.6	16	33.3	0.693	0.279	29	60.4	9	18.8
cosine overlap	0.622	0.276	31	64.6	6	12.5	0.618	0.265	28	58.3	4	8.3
KL inp-summ	-0.628	-0.262	28	58.3	8	16.7	-0.577	-0.267	22	45.8	6	12.5
topic overlap	0.597	0.265	30	62.5	5	10.4	0.689	0.277	26	54.2	3	6.3
% summ sign	0.607	0.269	23	47.9	7	14.6	0.534	0.272	23	47.9	1	2.1
mult. summ prob	0.434	0.268	8	16.7	0	0	0.459	0.272	10	20.8	0	0
uni. summ prob	0.292	0.261	2	4.2	0	0	0.466	0.287	2	4.2	0	0
regression	0.736	0.281	37	77.1	14	29.2	0.642	0.262	32	66.7	6	12.5
Rouge-1 recall	0.833	0.264	47	97.9	32	66.7	0.754	0.266	46	95.8	25	52.1
Rouge-2 recall	0.875	0.316	48	100	33	68.8	0.742	0.299	44	91.7	22	45.8

Table 3: Spearman correlations between feature values and manual system scores on a per input basis (TAC 2008 Query Focused summarization). Only the minimum, maximum values of the significant correlations are reported together with the number of significant correlations and the number of inputs for which correlations above 0.5 were obtained.

6 Evaluation of systems for TAC 2008 Update Summarization task

In the paper, we discussed only the results from evaluations of the query focused summaries produced at TAC 2008. The results for the update task are very similar and all conclusions hold for these data as well. For completeness we give the correlations between fully automatic and manual evaluations in Table 4.

7 Summarization as optimization—is JSD enough?

We have demonstrated that comparison of input and summary contents is predictive of summary quality. Further, our experiments show that a single best feature could approximate the comparison. A natural question arises in this situation—when a summarizer is built that globally optimizes for JS divergence during summary creation, wouldn't this input-based evaluation method be voided?

It must be remembered that the goal of summarization is not the selection of good content alone. Summarizers must also reduce redundancy, improve coherence and adhere to a length restriction. Often these goals might enter into conflicts during summary creation and satisfying them simultaneously becomes a difficult problem.

Studies of global inference algorithms for multi-document summarization (McDonald, 2007; tau Yih et al., 2007; Hassel and Sjöbergh, 2006) found that optimizing for content is NP-hard and equivalent to the Knapsack problem. (McDonald, 2007) further showed that intractability of a relevance maximization framework increases with the addition of redundancy constraints. Although, exact solutions may be found using ILP formulations, they can be used only on small document sets. Their huge runtimes makes them prohibitively expensive for summarizing large collections of documents. Hence only approximate solutions to the problem are feasible in real world situations.

Although some approximate solutions seem to obtain very good results in (McDonald, 2007), we must note that coherence is not included in that framework and that coherence is in fact a multi-faceted constraint requiring considerations of anaphors, discourse relations, cohesion and ordering. Together with coherence constraints, the inference could only become harder. Hence our evaluation method might still be suitable for content evaluation of summaries provided the overall summarizer scores include judgements of linguistic quality and redundancy as well.

8 Improving input-summary comparisons

Our experiments are clearly a starting point in understanding the role of inputs in summary evaluations. We demonstrated that simple comparison of summary and input contents using suitable features can capture perceptions of summary quality. These features can nevertheless be extended with other capabilities.

In fact, our test data comes from a query focused summarization task where a topic statement is also available for relevance assessment. We can expect better results by incorporating the query statement during evaluations. For example, we can select portions from the input that are relevant to the query and only use these for comparison with summary content.

For update summarization task, we experimented with different sets of features. Using averages of feature values comparing summary-background input and summary-update input, we obtained lower correlations with manual scores than when features based only on the update input were used. The summary- update in-

put features also outperform a linear regression metric which combines individual features from comparison with background and update inputs (Table 4). This result is not intuitive given the task definition. The background input is an important factor affecting the decision to include a particular content unit from the update set of documents. Further analysis needs to be carried out to ascertain the relative importance of the two input sets and how to best combine their features.

We also plan to expand our suite of features. A handful of other distributional similarity functions remain unexplored for our task and will be a readily accessible set of features— Euclidean distance, Jaccard's coefficient, L1 norm, confusion probability and skew divergence (Lee, 1999).

9 Conclusion

Summarization evaluation has always included human effort thereby limiting their scale and repeatability. In this paper, we have presented a successful framework for moving towards model-free evaluations—using the input as reference.

We have analyzed a variety of features for input/summary comparisons and demonstrated that the strength of different features vary, with certain features better suited for content comparisons. Low divergence from the input and diverse use of topic signatures in the summary are highly indicative of good content. We also find that preprocessing like stemming is useful in leveraging the capability of some features.

Very good results were obtained from a correlation analysis with human judgements, showing that input can indeed substitute for model summaries and manual efforts in summary evaluation. The best correlations were obtained by a single feature, JS divergence (0.9 with pyramid scores and 0.7 with responsiveness).

We have shown that the power of model-free evaluations generalizes across atleast two summarization tasks. Input is found useful in evaluating both query focused and update summaries. We have also presented a discussion on interesting questions on optimization and evaluation that arise as a result of this work and some future directions for input based evaluations.

References

Breck Baldwin, Robert Donaway, Eduard Hovy, Elizabeth Liddy, Inderjeet Mani, Daniel Marcu, Kathleen McKeown, Vibhu Mittal, Marc Moens, Dragomir Radev, Karen Sparck-Jones, Beth Sundheim, Simone Teufel, Ralph Weischedel, and Michael White. 2000. An Evaluation Road Map for Summarization Research. The Summarization Roadmap.

Ronald Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

John Conroy, Judith Schlesinger, and Dianne O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of ACL, short paper*.

Ido Dagan, Fernando Pereira, and Lillian Lee. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 272–278.

Robert L. Donaway, Kevin W. Drummey, and Laura A. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. In *NAACL-ANLP Workshop on Automatic Summarization*.

Donna Harman and Paul Over. 2004. The effects of human variation in duc summarization evaluation. In *ACL Text summarization branches out workshop*.

Martin Hassel and Jonas Sjöbergh. 2006. Towards holistic summarization: Selecting summaries, not sentences. In *Proceedings of LREC 2006*, Genoa, Italy.

Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *AAAI Symposium on Intelligent Summarization*.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI'05*.

Maria Lapata. 2000. The automatic interpretation of nominalizations. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 716–721.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32.

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

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurance statistics. In *Proceedings of HLT-NAACL 2003*.

Chin-Yew Lin, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 463–470.

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *ACL Text Summarization Workshop*.

Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, and Therese Firmin abd Beth Sundheim. 2002. Summac: a text summarization evaluation. *Natural Language Engineering*, 8(1):43–68.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *ECIR*, pages 557–564.

Kathleen McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Barry Schiffman, and Simone Teufel. 2001. Columbia multi-document summarization: Approach and evaluation. In *DUC'01*.

features	comparison with update inputs only pyramid score	responsiveness	avg. comparisons with update and background pyramid score	responsiveness
JSD divergence	-0.827	-0.764	-0.716	-0.669
JSD divergence smoothed	-0.825	-0.764	-0.713	-0.670
% of ip topic wds in summ	0.770	0.709	0.677	0.616
KL divergence summ-inp	-0.749	-0.709	-0.651	-0.624
KL divergence inp-summ	-0.741	-0.717	-0.644	-0.638
cosine inp-summ	0.727	0.691	0.649	0.631
% of summary = topic wd	0.721	0.707	0.647	0.636
topic overlap inp- summ	0.707	0.674	0.645	0.619
multinomial summ prob	0.284	0.355	0.152	0.224
unigram summ prob	-0.093	0.038	-0.151	-0.053
regression	0.789	0.605	0.699	0.522
regression combining features comparing with background and update inputs (without averaging) correlations = 0.8058 with pyramid 1, 0.6729 with responsiveness				

	pyramid score						responsiveness					
features	max	min	sig	%sig	a0.5	%a0.5	max	min	sig	%sig	a0.5	%a0.5
JSD smoothed	-0.753	-0.269	41	85.4	23	47.9	-0.747	-0.266	36	75.0	16	33.3
JSD	-0.746	-0.291	41	85.4	22	45.8	-0.738	-0.263	36	75.0	16	33.3
KL summ-inp	-0.739	-0.293	41	85.4	20	41.7	-0.705	-0.275	37	77.1	15	31.3
% of sign from inp	0.778	0.277	38	79.2	17	35.4	0.706	0.297	29	60.4	13	27.1
cosine overlap	0.665	0.275	33	68.8	10	20.8	0.685	0.267	28	14.6	7	14.6
% summ sign terms	0.737	0.263	32	66.7	11	22.9	0.672	0.265	28	58.3	6	12.5
topic overlap	0.679	0.264	31	64.6	9	18.8	0.665	0.274	26	54.2	5	10.4
KL inp-summ	-0.663	-0.281	30	62.5	9	18.8	-0.600	-0.285	24	50.0	5	10.4
mult. summ prob	0.479	0.267	12	25.0	0	0.0	0.547	0.262	13	27.1	1	2.1
uni. summ prob	0.363	0.362	1	2.1	0	0.0	0.266	0.266	1	2.1	0	0.0
regression	0.765	0.284	40	83.3	19	39.6	0.659	0.285	29	60.4	10	20.8
ROUGE-1 recall	0.842	0.392	48	100	41	85.4	0.811	0.268	46	95.8	30	62.5
ROUGE-2 recall	0.913	0.355	47	97.9	39	81.3	0.816	0.286	47	97.9	28	58.3

Table 4: Spearman correlations between fully automatic evaluation and manually assigned system scores for update summarization. Results are reported separately for features comparing update summaries with the update input only or with both update and background inputs and averaging the two (macro level). At the per-input level, only results for features comparing with update inputs are reported.

Kathleen McKeown, Rebecca Passonneau, David Elson, Ani Nenkova, and Julia Hirschberg. 2005. Do summaries help? a task-based evaluation of multi-document summarization. In *SIGIR*.

Andrew H. Morris, George M. Kasper, and Dennis A. Adams. 1992. The effects and limitations of automatic text condensing on reading comprehension. *Information System Research*, 3(1):17–35.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *HLT/NAACL*.

Ani Nenkova, Rebecca Passonneau, and Kathleen McKeown. 2007. The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Trans. Speech Lang. Process.*, 4(2):4.

Paul Over, Hoa Dang, and Donna Harman. 2007. Duc in context. *Inf. Process. Manage.*, 43(6):1506–1520.

Dragomir Radev and Daniel Tam. 2003. Single-document and multi-document summary evaluation via relative utility. In *Poster session, International Conference on Information and Knowledge Management (CIKM'03)*.

Dragomir Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale multi-document summarization: the mead project. In *Proceedings of ACL 2003*, Sapporo, Japan.

G. J. Rath, A. Resnick, and R. Savage. 1961. The formation of abstracts by the selection of sentences: Part 1: sentence selection by man and machines. *American Documentation*, 2(12):139–208.

Dmitri G. Roussinov and Hsinchun Chen. 2001. Information navigation on the web by clustering and summarizing query results. *Inf. Process. Manage.*, 37(6):789–816.

Tetsuya Sakai and Karen Sparck-Jones. 2001. Generic summaries for indexing in information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 190–198.

Ian Soboroff and Donna Harman. 2005. Novelty detection: the trec experience. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 105–112.

Wen tau Yih, Joshua Goodman, Lucy Vanderwende, and Hisami Suzuki. 2007. Multi-document summarization by maxi-

mizing informative content-words. In *Proceedings of IJCAI 2007*.

Anastasios Tombros and Mark Sanderson. 1998. Advantages of query biased summaries in information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–10.

Hans van Halteren and Simone Teufel. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *HLT-NAACL DUC Workshop*.