# Automatically Assessing Machine Summary Content Without a Gold Standard

Annie Louis*
University of Pennsylvania

Ani Nenkova**
University of Pennsylvania

*The most widely adopted approaches for evaluation of summary content follow some protocol for comparing a summary with gold-standard human summaries, which are traditionally called model summaries. This evaluation paradigm falls short when human summaries are not available and becomes less accurate when only a single model is available. We propose three novel evaluation techniques. Two of them are model-free and do not rely on a gold standard for the assessment. The third technique improves standard automatic evaluations by expanding the set of available model summaries with chosen system summaries.*

*We show that quantifying the similarity between the source text and its summary with appropriately chosen measures produces summary scores which replicate human assessments accurately. We also explore ways of increasing evaluation quality when only one human model summary is available as a gold standard. We introduce pseudomodels, which are system summaries deemed to contain good content according to automatic evaluation. Combining the pseudomodels with the single human model to form the gold-standard leads to higher correlations with human judgments compared to using only the one available model. Finally, we explore the feasibility of another measure—similarity between a system summary and the pool of all other system summaries for the same input. This method of comparison with the consensus of systems produces impressively accurate rankings of system summaries, achieving correlation with human rankings above 0.9.*

## 1. Introduction

In this work, we present evaluation metrics for summary content which make use of little or no human involvement. Evaluation methods such as manual pyramid scores (Nenkova, Passonneau, and McKeown 2007) and automatic ROUGE scores (Lin and Hovy 2003) rely on multiple human summaries as a gold standard (model) against which they compare a summary to assess how informative the candidate summary is. It is desirable that evaluation of similar quality be done quickly and cheaply

* E-mail: lannie@seas.upenn.edu.
** University of Pennsylvania, Department of Computer and Information Science, 3330 Walnut St., Philadelphia, PA 19104. E-mail: nenkova@seas.upenn.edu.

Submission received: 18 June 2011; revised submission received: 23 March 2012; accepted for publication: 18 April 2012.

doi:10.1162/COLI_a_00123

on non-standard test sets that have few or no human summaries, or on large test sets for which creating human model summaries is infeasible. In our work, we aim to identify indicators of summary content quality that do not make use of human summaries but can replicate scores based on comparison with a gold standard very accurately.

Such indicators would need to be easily computable from existing resources and to provide rankings of systems that agree with rankings obtained through human judgments. There have been some early proposals for alternative methods. Donaway, Drummey, and Mather (2000) propose that a comparison of the source text with a summary can tell us how good the summary is. A summary that has higher similarity with the source text can be considered better than one with lower similarity. Radev and Tam (2003) perform a large scale evaluation with thousands of test documents. Their work is set up in a search engine scenario. They first rank the test documents using the search engine. Then they perform the same experiment now substituting the summaries from one system in place of the original documents. The system whose summaries have the most similar ranking as that generated for the full documents is considered the best system because not much information loss is introduced by the summarization process.

But these methods did not gain much popularity and their performance was never compared to human evaluations. Part of the reason is that only in the last decade have several large data sets with system summaries and their ratings from human judges become available for performing such studies. Our work is the first to provide a comprehensive report of the strengths of such approaches and we show that human ratings can be reproduced by these fully automatic metrics with high accuracy. Our results are based on data for multi-document news summarization.

The key insights of our approach can be summarized as follows:

**Input–summary similarity:** Good summaries are representative of the input and so one would expect that the more similar a summary is to the input, the better its content. Identifying a suitable input–summary similarity metric will provide a means for fully automatic evaluation of summaries. We present a quantitative analysis of this hypothesis and show that input–summary similarity is highly predictive of scores assigned by humans for the summaries. The choice of an appropriate metric to measure similarity is critical, however, and we show that information-theoretic measures turn out to be the most powerful for this task (Section 4).

**Addition of pseudomodels:** Having a larger number of model summaries has been shown to give more stable evaluation results, but for some data sets only a single model summary is available. We test the utility of *pseudomodels*, which are system summaries that are chosen to be added to the human summary pool and that are used as additional models. We find that augmenting the gold standard with pseudomodels helps obtain better correlations with human judgments than if a single model is used (Section 5).

**System summaries as models:** Most current summarization systems perform content selection reasonably well. We examine an approach to evaluation that exploits system output and considers all system summaries for a given input as a gold standard (Section 6). We find that similarity between a summary and such a gold standard constitutes a powerful automatic evaluation measure. The correlation between this measure and human evaluations is over 0.9.

We analyze a number of similarity metrics to identify the ones that perform best for automatic evaluation. The tool we developed, SIMetrix (Summary Input similarity

Metrics), is freely available.[1] We test these resource-poor approaches to predict summary content scores assigned by human assessors. We evaluate the results on data from the Text Analysis Conferences.[2]

We find that our automatic methods to estimate summary quality are highly predictive of human judgments. Our best result is 0.93 correlation with human rankings using no model summaries and this is on par with automatic evaluation methods that do use human summaries. Our study provides some direction towards alternative methods of evaluation on non-standard test sets. The goal of our methods is to aid system development and tuning on new, especially large, data sets using little resources. Our metrics complement but are not intended to replace existing manual and automatic approaches to evaluation wherein the latter's strength and reliability are important for high confidence evaluations. Some of our findings are also relevant for system development as we identify desirable properties of automatic summaries that can be computed from the input (see Section 4). Our results are also strongly suggestive that system combination has the potential for improving current summarization systems (Section 6).

We start out with an outline of existing evaluation methods and the potential shortcomings of these approaches which we wish to address.

## 2. Current Content Evaluation Methods

Summary quality is defined by two key aspects—content and linguistic quality. A good summary should contain the most important content in the input and also structure the content and present it as well-written text. Several methods have been proposed for evaluating system-produced summaries; some only assess content, others only linguistic quality, and some combine assessment of both. Some of these approaches are manual and others can be performed automatically.

In our work, we consider the problem of automatic evaluation of content quality. To establish the context for our work, we provide an overview of current content evaluation methods used at the annual evaluations run by NIST.

The Text Analysis Conference (TAC, previously called the Document Understanding Conference [DUC][3]) conducts large scale evaluation of automatic systems on different summarization tasks. These conferences have been held every year since 2001 and the test sets and evaluation methods adopted by TAC/DUC have become the standard for reporting results in publications. TAC has employed a range of manual and automatic metrics over the years.

Manual evaluations of the systems are performed at NIST by trained assessors. The assessors score the summaries either

a) by comparing with a *gold-standard summary* written by humans, or

b) by providing a *direct rating* on a scale (1 to 5 or 1 to 10).

The human summaries against which other summaries are compared are interchangeably called models, gold standards, and references. Within TAC, they are typically called **models**.

1 SIMetrix can be downloaded at `http://www.seas.upenn.edu/~lannie/IEval2.html`.
2 `http://www.nist.gov/tac/`.
3 `http://duc.nist.gov/`.

### 2.1 Content Coverage Scores

The methods relying on a gold standard have evolved over the years. In the first years of DUC, a single model summary was used. System summaries were evaluated by manually assessing how much of the model's content is expressed in the system summary. Each clause in the model represents one unit for the evaluation. For each of these clauses, assessors specify the extent to which its content is expressed in a given system summary. The average degree to which the model summary's clauses overlap with the system summary's content is called **coverage**. These coverage scores were taken as indicators of content quality for the system summaries.

Different people include very different content in their summaries, however, and so the coverage scores can vary depending on which model is used (Rath, Resnick, and Savage 1961). This problem of bias in evaluation was later addressed by the pyramid technique, which combines information from multiple model summaries to compose the reference for evaluation. Since 2005, the pyramid evaluation method has become standard.

### 2.2 Pyramid Evaluation

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, Passonneau, and McKeown 2007). It uses multiple human models from which annotators identify semantically defined Summary Content Units (SCUs). 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. The pyramid score for a system summary $S$ is equal to the following ratio:

$$py(S) = \frac{\text{sum of weights of SCUs expressed in } S}{\text{sum of weights of an ideal summary with the same number of SCUs as } S} \quad (1)$$

In this way, a more reliable score for a summary is obtained using multiple reference summaries. Four human summaries are normally used for pyramid evaluation at TAC.

### 2.3 Responsiveness Evaluation

**Responsiveness** of a summary is a measure of overall quality combining both content selection and linguistic quality. It measures to what extent summaries convey appropriate content in a structured fashion. Responsiveness is assessed by direct ratings given by the judges. For example, a scale of 1 (poor summary) to 5 (very good summary) is used and these assessments are done without reference to any model summaries.

Pyramid and responsiveness are the standardly used manual approaches for content evaluation. They produce rather similar rankings of systems at TAC. The (Spearman) correlation between the two for ranking systems that participated in the TAC 2009 conference is 0.85 (p-value 6.8e-16, 53 systems). The responsiveness measure involves some aspects of linguistic quality whereas the pyramid metric was designed for content only. Such high correlation indicates that the content factor has

substantial influence on the responsiveness judgments, however. The high correlation also indicates that two types of human judgments made on very different basis—gold-standard summaries and direct judgments—can agree and provide fairly similar rankings of summaries.

### 2.4 ROUGE

Manual evaluation methods require significant human effort. Moreover, the pyramid evaluation involves detailed annotation for identifying SCUs in human and system summaries and requires training of assessors to perform the evaluation. Outside of TAC, therefore, system developments and results are regularly reported using ROUGE, a suite of automatic evaluation metrics (Lin and Hovy 2003; Lin 2004b).

ROUGE automates the comparison between model and system summaries based on *n*-gram overlaps. These overlap scores have been shown to correlate well with human assessment (Lin 2004b) and so ROUGE removes the need for manual judgments in this part of the evaluation.

ROUGE scores are computed typically using unigram (R1) or bigram (R2) overlaps. In TAC, four human summaries are used as models and their contents are combined for computing the overlap scores. For fixed length summaries, the recall from the comparison is used as the quality metric. Other metrics such as longest subsequence match are also available. Another ROUGE variant is RSU4, which computes the overlap in terms of **skip bigrams**, where two unigrams with a gap of up to four intervening words are considered as bigrams. This latter metric provides some additional flexibility compared to the stricter R2 scores.

The correlations between ROUGE and manual evaluations for systems in TAC 2009 are shown in Table 1 and vary between 0.76 and 0.94 for the different variants.[4] Here, and in all subsequent experiments, Spearman correlations are computed using the R toolkit (R Development Core Team 2011). In this implementation, significance values for the correlations are produced using the AS 89 algorithm (Best and Roberts 1975).

These correlations are highly significant and show that ROUGE is a high performance automatic evaluation metric.

We can consider the ROUGE results as the upper bound of performance for the model-free evaluations that we propose because ROUGE involves direct comparison with the gold-standard summaries. Our metrics are designed to be used when model summaries are not available.

### 2.5 Automatic Evaluation Without Gold-Standard Summaries

All of these methods require significant human involvement. In evaluations where gold-standard summaries are needed, assessors first read the input documents (10 or more per input) and write a summary. Then manual comparison of system and gold standard is done, which takes additional time. Gillick and Liu (2010) hypothesize that at least 17.5 hours are needed to evaluate two systems under this set up on a standard test set. Moreover, multiple gold-standard summaries are needed for the same input, so different assessors have to read and create summaries. The more reliable evaluation

4 The scores were computed after stemming but stop words were retained in the summaries.

**Table 1**
Spearman correlation between manual scores and ROUGE metrics on TAC 2009 data (53 systems). All correlations are highly significant with p-value $< 10^{-10}$.

| ROUGE variant | Pyramid | Responsiveness |
|---|---|---|
| ROUGE-1 | 0.88 | 0.76 |
| ROUGE-2 | 0.94 | 0.82 |
| ROUGE-SU4 | 0.92 | 0.79 |

methods such as pyramid involve even more annotations at the clause level. Although responsiveness does not require gold-standard summaries, in a system development setting, responsiveness judgments are resource-intensive. It requires judges to directly assign scores to summaries, so humans are in the loop each time the evaluation needs to be done, making it rather costly. For ROUGE, however, once the human summaries are created, the scores can be computed automatically for repeated system development runs. This benefit has made ROUGE immensely popular. But the initial investment of time for gold-standard creation is still necessary.

Another important point is that for TAC, the gold standards are created by trained assessors at NIST. Non-expert evaluation options such as Mechanical Turk have recently been explored by Gillick and Liu (2010). They provided annotators with gold-standard references and system summaries and asked them to score the system summaries on a scale from 1 to 10 with respect to how well they convey the same information as the models. They analyzed how these scores are related to responsiveness judgments given by the expert TAC assessors. The study assessed only eight automatic systems from TAC 2009 and the correlation between the ratings from experts and Mechanical Turk annotations was 0.62 (Spearman). The analysis concludes that evaluations produced in this way tend to be noisy.

One reason was that non-expert annotators were quite influenced by the readability of the summaries. For example, they tended to assign high scores to the baseline summary that picks the lead paragraph. The baseline summary, however, is ranked by expert annotators as low in responsiveness compared to other systems' summaries. Further, the non-expert evaluation led to few significant differences in the system rankings (score of system A is significantly greater/lesser than that of B) compared with the TAC evaluations of the same systems.

Another problem with non-expert evaluation is the quality of the model summaries. Evaluations based on model summaries assume that the gold standards are of high quality. Through the years at TAC, considerable effort has been invested to ensure that the evaluation scores do not vary depending on the particular gold standard. In the early years of TAC only one gold-standard summary was used. During this time, papers reported ANOVA tests examining the factors that most influenced summary scores from the evaluations and found that the identity of the judge turned out to be the most significant factor (McKeown et al. 2001; Harman and Over 2004). But it is desirable that a model summary or a human judgment be representative of important content in general and does not depict the individual biases of the person who created the summary or made the judgment. So the evaluation methodology was refined to remove the influence of the assessor identity on the evaluation. The pyramid evaluation was also developed with this goal of smoothing out the variation between judges. Gillick and Liu (2010) point out that Mechanical Turk evaluations have this undesirable outcome: The identity

of the judges turns out to be the most significant factor influencing summary scores. Gillick and Liu do not elicit model summaries, only direct judgments on quality. We suspect that the task would only be harder if model summaries were to be created by non-experts.

The problem that has been little addressed by any of these discussed metrics is evaluation when there are no gold-standard summaries available. Systems are developed by fine-tuning on the TAC data sets, but in non-TAC data sets in novel or very large domains model summaries may not be available. Even though ROUGE provides good performance in automatic evaluation, it is not usable under these conditions. Further, pyramid and ROUGE use multiple gold-standard summaries for evaluation (ROUGE correlates with human judgments better when computed using multiple models; we discuss this aspect further in Section 5) so even a single gold-standard summary may not be sufficient for reliable evaluation.

In our work, we propose fully automatic methods for content evaluation which can be used in the absence of human summaries. We also explore methods to further improve the evaluation performance when only one model summary is available.

## 3. Data and Evaluation Plan

In this section, we describe the data we use throughout our article. We carry out our analysis on the test sets and system scores from TAC 2009. TAC 2009 is also the year when NIST introduced a special track called AESOP (Automatically Evaluating Summaries of Peers). The goal of AESOP is to identify automatic metrics that correlate well with human judgments of summary quality.

We use the data from the TAC 2009 query focused-summarization task.[5] Each input consists of ten news documents. In addition, the user's information needs associated with each input is given by a query statement consisting of a title and narrative. An example query statement is shown here:

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

A system must produce a summary that addresses the information required by the query. The maximum length for summaries is 100 words.

The test set contains 44 inputs, and 53 automatic systems (including baselines) participated that year. These systems were manually evaluated for content using both pyramid and responsiveness methods. In TAC 2009, two oracle systems were introduced during evaluation whose outputs are in fact summaries created by people. We ignore these two systems and use only the automatic participant submissions and the automatic baseline systems.

As a development set, we use the inputs, summaries, and evaluations from the previous year, TAC 2008. There were 48 inputs in the query-focused task in 2008 and 58 automatic systems participated.

TAC 2009 also involved an update summarization task and we obtained similar results on the summaries from this task. In this article, for clarity we only present results

5 http://www.nist.gov/tac/2009/Summarization/update.summ.09.guidelines.html.

on evaluating the query-focused summaries, but the update task results are described in detail in Louis and Nenkova (2008, 2009a, 2009c).

### 3.1 Evaluating Automatic Metrics

For each of our proposed metrics, we need to assess their performance in replicating manually produced rankings given by the pyramid and responsiveness evaluations. We use two measures to compare these human scores for a system with the automatic scores from one of our metrics:

a) SPEARMAN CORRELATION: Reporting correlations with human evaluation metrics is the norm for validating automatic metrics. We report Spearman correlation, which compares the rankings of systems produced by the two methods instead of the actual scores assigned to systems.

b) PAIRWISE ACCURACY: To complement correlation results with numbers that have easier intuitive interpretation, we also report the pairwise accuracy of our metrics in predicting the human scores. For every pair of systems ($A$, $B$), we examine whether their pairwise ranking (either $A > B$, $A < B$, or $A = B$) according to the automatic metric agrees with the ranking of the same pair according to human evaluation. If it does, the pair is **concordant** with human judgments. The pairwise accuracy is the percentage of concordant pairs out of the total system pairs. This accuracy measure is more interpretable than correlations in terms of the errors made by a metric. A metric with 90% accuracy incorrectly flips 10% of the pairs, on average, in a ranking it produces. This measure is inspired by the Kendall tau coefficient.

We test the metrics for success in replicating human scores overall across the full test set as well as identifying good and bad summaries for individual inputs. We therefore report the correlation and accuracy of our metrics at the following two levels.

a) SYSTEM LEVEL (MACRO): The average score for a system is computed over the entire set of test inputs using both manual and our automatic methods. The correlations between ranks assigned to systems by these average scores will be indicative of the strength of our features to predict overall system rankings on the test set. Similarly, the pairwise accuracies are computed using the average scores for the systems in the pair.

b) INPUT LEVEL (MICRO): For each individual input, we compare the rankings for the system summaries using manual and automatic evaluations. Here the correlation or accuracy is computed for each input. For correlations, we report the percentage of inputs for which significant correlations (p-value $< 0.05$) were obtained. For accuracy, the systems are paired within each input. Then these pairs for all the inputs are put together and the fraction of concordant pairs is computed. Micro-level analysis highlights the ability of an evaluation metric to identify good and poor quality system summaries produced for a specific input and this task is bound to be harder than system level predictions. For example, even with wrong prediction of rankings on a few inputs, the average scores (macro-level) for a system might not be affected.

In the following sections, we describe three experiments in which we analyze the possibility of performing automatic evaluation involving only minimal or no human judgments: Using input–summary similarity (Section 4), using system summaries as pseudomodels alongside gold-standard summaries created by people (Section 5), and using the collection of system summaries as a gold standard (Section 6). All the automatic systems, including baselines, were evaluated.

## 4. Input–Summary Similarity: Evaluation Using Only the Source Text

Here we present and evaluate a suite of metrics which do not require gold-standard human summaries for evaluation. The underlying intuition is that good summaries will tend to be similar to the input in terms of content. Accordingly, we use the similarity of the distribution of terms in the input and summaries as a measure of summary content.

Although the motivation for this metric is highly intuitive, it is not clear how similarity should be defined for this particular problem. Here we provide a comprehensive study of input–summary similarity metrics and show that some of these measures can indeed be very accurate predictors of summary quality even while using no gold-standard human summaries at all.

Prior to our work, the proposal for using the input for evaluation has been brought up in a few studies. These studies did not involve a direct evaluation of the capacity of input–summary similarity to replicate human ratings, however, and they did not compare similarity metrics for the task. Because large scale manual evaluation results are available now, our work is the first to evaluate this possibility in a direct manner and involving study of correlations with different types of human evaluations. In the following section we detail some of the prior studies on input–summary similarity for summary evaluation.

### 4.1 Related Work

One of the motivations for using the input text rather than gold-standard summaries comes from the need to perform large scale evaluations with test sets comprised of thousands of inputs. Creating human summaries for all of them would be an impossible task indeed.

In Radev and Tam (2003), therefore, a large scale fully automatic evaluation of eight summarization systems on 18,000 documents was performed without any human effort by using the idea of input–summary similarity. A search engine was used to rank documents according to their relevance to a given query. The summaries for each document were also ranked for relevance with respect to the same query. For good summarization systems, the relevance ranking of summaries is expected to be similar to that of the full documents. Based on this intuition, the correlation between relevance rankings of summaries and original documents was used to compare the different systems. A system whose summaries obtained highly similar rankings to the original documents can be considered better than a system whose rankings have little agreement.

Another situation where input–summary similarity was hypothesized as a possible evaluation was in work concerned with reducing human bias in evaluation. Because humans vary considerably in the content they include for the same input (Rath, Resnick, and Savage 1961; van Halteren and Teufel 2003), rankings of systems are rather different depending on the identity of the model summary used (also noted by McKeown et al. [2001] and Jing et al. [1998]). Donaway, Drummey, and Mather (2000) therefore suggested that there are considerable benefits to be had in adopting a method of evaluation that does not require human gold standards but instead directly compares the original document and its summary. In their experiments, Donaway, Drummey, and Mather demonstrated that the correlations between manual evaluation using a gold-standard summary and

a) manual evaluation using a different gold-standard summary

b) automatic evaluation by directly comparing input and summary[6]

are the same. Their conclusion was that such automatic methods should be seriously considered as an alternative to evaluation protocols built around the need to compare with a gold standard.

These studies, however, do not directly assess the performance of input–summary similarity for ranking systems. In Louis and Nenkova (2009a), we provided the first study of several metrics for measuring similarity for this task and presented correlations of these metrics with human produced rankings of systems. We have released a tool, SIMetrix (Summary-Input Similarity Metrics), which computes all the similarity metrics that we explored.[7]

### 4.2 Metrics for Computing Similarity

In this section, we describe a suite of similarity metrics for comparing the input and summary content. We use cosine similarity, which is standard for many applications. The other metrics fall under three main classes: distribution similarity, summary likelihood, and use of topic signature words. The distribution similarity metrics compare the distribution of words in the input with those in the summary. The summary likelihood metrics are based on a generative model of word probabilities in the input and use the model to compute the likelihood of the summary. Topic signature metrics focus on a small set of descriptive and topical words from the input and compare them to summary content rather than using the full vocabulary of the input.

Both input and summary words were stopword-filtered and stemmed before computing the features.

*4.2.1 Distribution Similarity.* Measures of similarity between two probability distributions are a natural choice for our task. One would expect good summaries to be characterized by low divergence between probability distributions of words in the input and summary, and by high similarity with the input.

We experimented with three common measures: Kullback Leibler (KL) divergence, Jensen Shannon (JS) divergence, and cosine similarity.

These three metrics have already been applied for summary evaluation, albeit in a different context. In their study of model-based evaluation, Lin et al. (2006) used KL and JS divergences to measure the similarity between human and machine summaries. They found that JS divergence always outperformed KL divergence. Moreover, the performance of JS divergence was better than standard ROUGE scores for multi-document summarization when multiple human models were used for the comparison.

The use of input–summary similarity in Donaway, Drummey, and Mather (2000), which we described in the previous section, is more directly related to our work. But here, inputs and summaries were compared using only one metric: cosine similarity.

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

$$D(P||Q) = \sum_{w} p_P(w) \log_2 \frac{p_P(w)}{p_Q(w)} \tag{2}$$

6 They used cosine similarity to perform the input–summary comparison.

7 `http://www.seas.upenn.edu/~lannie/IEval2.html`.

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 of word probabilities are estimated from the input and summary, respectively. Because KL divergence is not symmetric, both input–summary and summary–input divergences are introduced as metrics. In addition, the divergence is undefined when $p_P(w) > 0$ but $p_Q(w) = 0$. We perform simple smoothing to overcome the problem.

$$p(w) = \frac{C + \delta}{N + \delta * B} \tag{3}$$

Here $C$ is the count of word $w$ and $N$ is the number of tokens; $B = 1.5|V|$, where $V$ is the input vocabulary and $\delta$ 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 incorporates 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

$$J(P||Q) = \frac{1}{2}[D(P||A) + D(Q||A)], \tag{4}$$

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 compute both smoothed and unsmoothed versions of the divergence as summary scores.

**Vector space similarity:** The third metric is cosine overlap between the $tf * idf$ vector representations of input and summary contents.

$$cos\theta = \frac{v_{inp} \cdot v_{summ}}{||v_{inp}||\ ||v_{summ}||} \tag{5}$$

We compute two variants:

1. Vectors contain all words from input and summary.
2. Vectors contain only topic signature words from the input and all words of the summary.

Topic signatures are words highly descriptive of the input, as determined by the application of the log-likelihood test (Lin and Hovy 2000). Using only topic signatures from the input to represent text is expected to be more accurate because the reduced vector has fewer dimensions compared with using all the words from the input.

*4.2.2 Summary Likelihood.* For this approach, we view summaries as being generated according to word distributions in the input. Then the probability of a word in the input would be indicative of how likely it is to be emitted into a summary. Under this generative model, the likelihood of a summary's content can be computed using different methods and we expect the likelihood to be higher for better quality summaries.

We compute both a summary's unigram probability as well as its probability under a multinomial model.

**Unigram summary probability:**

$$(p_{inp}w_1)^{n_1}(p_{inp}w_2)^{n_2}...(p_{inp}w_r)^{n_r} \quad (6)$$

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 $w_1 \dots w_r$ are all words in the summary vocabulary.

**Multinomial summary probability:**

$$\frac{N!}{n_1!n_2!\dots n_r!}(p_{inp}w_1)^{n_1}(p_{inp}w_2)^{n_2}\dots(p_{inp}w_r)^{n_r} \quad (7)$$

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

*4.2.3 Use of Topic Words in the Summary.* Summarization systems that directly optimize the number of topic signature words during content selection have fared very well in evaluations (Conroy, Schlesinger, and O'Leary 2006). Hence the number of topic signatures from the input present in a summary might be a good indicator of summary content quality. In contrast to the previous methods, by limiting to topic words, we use only a representative subset of the input's words for comparing with summary content.

We experiment with two features that quantify the presence of topic signatures in a summary:

1. The fraction of the summary composed of input's topic signatures.
2. The percentage of topic signatures from the input that also appear in the summary.

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

*4.2.4 Feature Combination Using Linear Regression.* We also evaluated the performance of a linear regression metric combining all of these features. During development, the value of the regression-based score for each summary was obtained using a leave-one-out approach. For a particular input and system-summary combination, 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.

### 4.3 Results

We first present an analysis of all the similarity metrics on our development data, TAC'08. In the next section, we analyze the performance of our two best features on the TAC'09 data set.

*4.3.1 Feature Analysis: Which Similarity Metric is Best?* Table 2 shows the *macro-level* Spearman correlations between manual and automatic scores averaged across the 48 inputs in TAC'08.

Overall, we find that both distribution similarity and topic signature features produce system rankings very similar to those produced by humans. Summary likelihood, on the other hand, turns out to not be predictive of content selection performance. The

**Table 2**
Spearman correlation on the macro level for TAC'08 data (58 systems). All results are highly significant with p-values $< 0.000001$ except unigram and multinomial summary probability, which are not significant even at the 0.05 level.

| Features | Pyramid | Responsiveness |
|---|---|---|
| JS div | −0.880 | −0.736 |
| JS div smoothed | −0.874 | −0.737 |
| % of input topic words | 0.795 | 0.627 |
| KL div summary–input | −0.763 | −0.694 |
| cosine overlap, all words | 0.712 | 0.647 |
| % of summary = topic words | 0.712 | 0.602 |
| cosine overlap, topic words | 0.699 | 0.629 |
| KL div input–summary | −0.688 | −0.585 |
| multinomial summary probability | 0.222 | 0.235 |
| unigram summary probability | −0.188 | −0.101 |
| regression | 0.867 | 0.705 |
| | | |
| ROUGE-1 recall | 0.859 | 0.806 |
| ROUGE-2 recall | 0.905 | 0.873 |

linear regression combination of features obtains high correlations with manual scores but does not lead to better results than the single best feature: JS divergence.

JS divergence obtains the best correlations with both types of manual scores—0.88 with pyramid score and 0.74 with responsiveness. The regression metric performs comparably, with correlations of 0.86 and 0.70. The correlations obtained by both JS divergence and the regression metric with pyramid evaluations are in fact better than that obtained by ROUGE-1 recall (0.85).

The best topic signature-based feature—the percentage of input's topic signatures that are present in the summary—ranks next only to JS divergence and regression. The correlations between this feature and pyramid and responsiveness evaluations are 0.79 and 0.62, respectively. The proportion of summary content composed of topic words performs worse as an evaluation metric with correlations 0.71 and 0.60. This result indicates that summaries that cover more topics from the input are judged to have better content than those in which fewer topics are mentioned.

Cosine overlaps and KL divergences obtain good correlations but still lower than JS divergence and the percentage of input topic words. Further, rankings based on unigram and multinomial summary likelihood do not correlate significantly with manual scores.

On a per input basis, the proposed metrics are not that effective in distinguishing which summaries have good and poor content. The minimum and maximum correlations with manual evaluations across the 48 inputs are given in Table 3. The number and percentage of inputs for which correlations were significant are also reported.

JS divergence obtains significant correlations with pyramid scores for 73%. The best correlation was 0.71 on a particular input and the worst performance was 0.27 correlation for another input. The results are worse for other features and for comparison with responsiveness scores.

At the micro level, combining features with regression gives the best result overall, in contrast to the findings for the macro-level setting. This result has implications for system development; no single feature can reliably predict good content for a particular input. Even a regression combination of all features is a significant predictor of

**Table 3**
Spearman correlations at micro level for TAC'08 data (58 systems). Only the minimum and maximum values of the significant correlations are reported, together with the number and percentage of inputs that obtained significant correlation.

| | Pyramid | | | Responsiveness | | |
|---|---|---|---|---|---|---|
| **Features** | **max** | **min** | **number significant (%)** | **max** | **min** | **number significant (%)** |
| JS div | −0.714 | −0.271 | 35 (72.9) | −0.654 | −0.262 | 35 (72.9) |
| JS div smoothed | −0.712 | −0.269 | 35 (72.9) | −0.649 | −0.279 | 33 (68.8) |
| KL div summary-input | −0.736 | −0.276 | 35 (72.9) | −0.628 | −0.261 | 35 (72.9) |
| % of input topic words | 0.701 | 0.286 | 31 (64.6) | 0.693 | 0.279 | 29 (60.4) |
| cosine overlap - all words | 0.622 | 0.276 | 31 (64.6) | 0.618 | 0.265 | 28 (58.3) |
| KL div input-summary | −0.628 | −0.262 | 28 (58.3) | −0.577 | −0.267 | 22 (45.8) |
| cosine overlap - topic words | 0.597 | 0.265 | 30 (62.5) | 0.689 | 0.277 | 26 (54.2) |
| % summary = topic words | 0.607 | 0.269 | 23 (47.9) | 0.534 | 0.272 | 23 (47.9) |
| multinomial summary prob. | 0.434 | 0.268 | 8 (16.7) | 0.459 | 0.272 | 10 (20.8) |
| unigram summary prob. | 0.292 | 0.261 | 2 (4.2) | 0.466 | 0.287 | 2 (4.2) |
| regression | 0.736 | 0.281 | 37 (77.1) | 0.642 | 0.262 | 32 (66.7) |
| ROUGE-1 recall | 0.833 | 0.264 | 47 (97.9) | 0.754 | 0.266 | 46 (95.8) |
| ROUGE-2 recall | 0.875 | 0.316 | 48 (100) | 0.742 | 0.299 | 44 (91.7) |

content selection quality in only 77% of the cases. For example, a set of documents, each describing a different opinion on an issue, is likely to have less repetition on both the lexical and content unit levels. Because the input–summary similarity metrics rely on the word distribution of the input for clues about important content, their predictiveness will be limited for such inputs.[8] Follow-up work to our first results on fully automatic evaluation by Saggion et al. (2010) has assessed the usefulness of the JS divergence measure for evaluating summaries from other tasks and for languages other than English. Whereas JS divergence was significantly predictive of summary quality for other languages as well, it did not work well for tasks where opinion and biographical type inputs were summarized. We provide further analysis and some examples in Section 7.

Overall, the micro level results suggest that the fully automatic measures we examined will not be useful for providing information about summary quality for an individual input. For averages over many test sets, the fully automatic evaluations give more reliable results, and are highly correlated with rankings produced by manual evaluations. On the other hand, model summaries written for the specific input would give a better indication of what information in the input was important and interesting. This is indeed the case as we shall see from the ROUGE scores in the next section.

*4.3.2 Comparison with ROUGE.* The aim of our study is to assess metrics for evaluation in the absence of human gold standards, scenarios where ROUGE cannot be used. We do not intend to directly compare the performance of ROUGE with our metrics,

8 In fact, it would be surprising to find an automatically computable feature or feature combination which would be able to consistently predict good content for all individual inputs. If such features existed, an ideal summarization system would already exist.

therefore. We discuss the correlations obtained by ROUGE in the following, however, to provide an idea of the reliability of our metrics compared with evaluation quality that is provided by ROUGE and multiple human summaries.

At the macro level, the correlation between ROUGE-1 and pyramid scores is 0.85 (Table 2). For ROUGE-2 the correlation with pyramid scores is 0.90, practically identical with JS divergence.

Because the performance of these two measures seem close, we further analyzed their errors. The focus of this analysis is to understand if JS divergence and ROUGE-2 are making errors in ordering the same systems or whether their errors are different. This result would also help us to understand if ROUGE and JS divergence have complementary strengths that can be combined. For this, we considered pairs of systems and computed the better system in each pair according to the pyramid scores. Then, for ROUGE-2 and JS divergence, we recorded how often they provided the correct judgment for the pairs as indicated by the pyramid evaluation. There were 1,653 pairs of systems at the macro level and the results are in Table 4.

This table shows that a large majority (80%) of the same pairs are correctly predicted by both ROUGE and JS divergence. Another 6% of the pairs are such that both metrics do not provide the correct judgment. Therefore, ROUGE and JS divergence appear to agree on a large majority of the system pairs. There is a small percentage (14%) that is correctly predicted by only one of the metrics. The chances of combining ROUGE and JS divergence to get a better metric appears small, therefore. To test this hypothesis, we trained a simple linear regression model combining JS divergence and ROUGE-2 scores as predictors for the pyramid scores and tested the predictions of this model on data from TAC 2009. The combination did not give improved correlations compared with using ROUGE-2 alone.

In the case of manual responsiveness, which combines aspects of linguistic quality along with content selection evaluation, the correlation with JS divergence is 0.73. For ROUGE, it is 0.80 for R1 and 0.87 for R2. Here, ROUGE-1 outperforms all the fully automatic evaluations. This is evidence that the human gold-standard summaries provide information that is unlikely to ever be approximated by information from the input alone, regardless of feature sophistication.

At the micro level, ROUGE clearly does better than all the fully automatic measures for replicating both pyramid and responsiveness scores. The results are shown in the last two rows of Table 3. ROUGE-1 recall obtains significant correlations for over 95% of inputs for responsiveness and 98% of inputs for pyramid evaluation compared to 73% (JS divergence) and 77% (regression). Undoubtedly, at the input level, comparison with model summaries is substantially more informative.

When gold-standard summaries are not available, however, our features can provide reliable estimates of system quality when averaged over a set of test inputs.

**Table 4**
Overlap between ROUGE-2 and JS divergence predictions for the best system in a pair (TAC 2008, 1,653 pairs). The gold-standard judgment for a better system is computed using the pyramid scores.

| | **JSD correct** | **JSD incorrect** |
|---|---|---|
| **ROUGE-2 correct** | 1,319 (79.8%) | 133 (8.1%) |
| **ROUGE-2 incorrect** | 96 (5.8%) | 105 (6.3%) |

**Table 5**
Input–summary similarity evaluation: Results on TAC'09 (53 systems).

| | Correlations | | | | Pairwise accuracy | | | |
|---|---|---|---|---|---|---|---|---|
| | Macro level | | Micro level | | Macro level | | Micro level | |
| **Metric** | **py** | **resp** | **py** | **resp** | **py** | **resp** | **py** | **resp** |
| JS div | 0.74 | 0.70 | 84.1 | 75.0 | 78.0 | 75.7 | 65.1 | 50.1 |
| Regr | 0.77 | 0.67 | 81.8 | 65.9 | 80.1 | 74.8 | 64.7 | 49.4 |
| RSU4 - 4 models | 0.92 | 0.79 | 95.4 | 81.8 | 88.4 | 80.0 | 70.5 | 53.0 |

*4.3.3 Results on TAC'09 Data.* To evaluate our metrics for fully automatic evaluation, we make use of the TAC'09 data. The regression metric was trained on all of the 2008 data with pyramid scores as the target. Table 5 shows the results on the TAC'09 data. We also report the correlations obtained by ROUGE-SU4 because it was the official baseline measure adopted at TAC'09 for comparison of automatic evaluation metrics.

The correlations are lower than on our development set. The highest correlation at macro level is 0.77 (regression) in contrast to 0.88 (JS divergence) and 0.86 (regression) obtained on the TAC'08. The regression metric turns out better than JS divergence on the TAC'09 data for predicting pyramid scores. JS divergence continues to be the best metric on the basis of correlations with responsiveness, however.

In terms of the pairwise scores, the automatic metrics have 80% accuracy in predicting the pyramid scores at the system level, about 8% lower than that obtained by ROUGE. For responsiveness, the best accuracy is obtained by regression (75%). This result shows that the ranking according to responsiveness is likely to have a large number of flips. ROUGE is 5 percentage points better than regression for predicting responsiveness but this value is still low compared to accuracies in replicating the pyramid scores.

The pairwise accuracy at the micro level is 65% for the automatic metrics and here the gap between ROUGE and our metrics is 5 percentage points but it is a significant percentage as the total pairs at micro level are about 60,000 (all pairings of 53 systems in 44 inputs).

Overall, the performance of the fully automatic evaluation is still high for use during system development. A further advantage is that these metrics are consistently predictive across two years as shown by these results. In Section 7, we analyze some reasons for the difference in performance in the two years. In terms of best metrics, both JS divergence and regression turn out to be useful with little difference in performance between them.

## 5. Pseudomodels: Use of System Summaries in Addition to Human Summaries

Methods such as pyramid use multiple human summaries to avoid bias in evaluation when using a single gold standard. ROUGE metrics are also currently used with multiple models, when available. But often, even if gold-standard summaries are available on non-standard test sets, they are few in number. Data sets with one gold-standard summary (such as abstracts of scientific papers and editor-produced summaries of news articles) are common. The question now is whether we can provide the same quality

evaluation using a single gold-standard summary as compared to using several gold standards.

To tackle this problem, we propose the use of pseudomodel system summaries. Our approach is as follows: We first predict the scores of systems on the basis of the few available models. The top ranking systems from this evaluation are then considered as "pseudo-models;" their summaries are added to the gold-standard set along with the existing human models. The final evaluation scores are produced by comparison with this expanded model set—original model summaries plus the pseudomodels. Our hypothesis is that the scores produced after the addition of pseudomodels would be more reliable and correlate better with human scores compared with evaluation using a single model summary.

Before we describe our method, we provide a glimpse of the variation in evaluation quality depending on the number of models used. Previous studies have shown that at the system level, system rankings even with a single model will be stable when computed over a large enough number of test inputs. Harman and Over (2004) show that the relative ranks of systems computed using one model do not change when computed using another model when the number of inputs is large. Again under the same conditions of having a large number of inputs, Lin (2004a) and Owkzarzak and Dang (2009) show that ROUGE correlations with human scores are stable when using few human models. In machine translation evaluation, similar results are noted by Zhang and Vogel (2010), who found that the lack of additional reference translations can be handled by evaluating the systems on more test examples.

Multiple models are particularly important for evaluation at the level of individual inputs, however. Table 6 shows the difference in correlations and pairwise accuracy of ROUGE with human scores when one and four model summaries are used. We picked the first model in alphabetical order of their names for the computation of correlation between metrics and a single model.

At the system level, the correlations from both set-ups are similar. But at the micro level, there is considerable difference in performance. Using all four models, significant correlations with pyramid scores are obtained for 95% of the inputs. The evaluations that rely on a single model produce significant correlations for only 84% of the inputs, however. For responsiveness scores, which are model-independent, we see that the micro-level evaluations have a smaller increase as more models are added (79% to 81%). Again in terms of pairwise accuracy, the accuracy in predicting micro-level pyramid scores improves by 4% when additional models are used and the improvement is 3% for predicting responsiveness scores. Given this difference in performance when one and many models are used, we investigate how to improve evaluation when only one model is available.

**Table 6**
ROUGE evaluation with different number of models: macro level (Spearman correlations), micro level (percentage of inputs with significant correlations on TAC'09 data). No. of systems = 53.

| | **Correlations** | | | | **Pairwise accuracy** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Macro level** | | **Micro level** | | **Macro level** | | **Micro level** | |
| **Task** | **py** | **resp** | **py** | **resp** | **py** | **resp** | **py** | **resp** |
| RSU4 - 1 model | 0.92 | 0.80 | 84.1 | 79.5 | 88.3 | 80.3 | 66.1 | 50.7 |
| RSU4 - 4 models | 0.92 | 0.79 | **95.4** | **81.8** | 88.4 | 80.0 | **70.5** | **53.0** |

We explore the possibility of augmenting the model set with *good* system summaries. These system summaries or "pseudomodels" are chosen to be the ones which receive high scores based on the one available model summary. We expect that the benefit of pseudomodels will be noticeable in micro-level correlations with pyramid scores. At the macro level, even with multiple human models there is no improvement in correlations compared with a single model, and the addition of less-ideal system summaries is not likely to be better than adding human summaries.

### 5.1 Related Work

The idea of using system output for evaluation was introduced in the context of machine translation by Albrecht and Hwa (2007, 2008). In their method, Albrecht and Hwa (2007) designate some systems to act as pseudoreferences. Then, every candidate translation to be evaluated is compared to the translations produced by the pseudoreferences using a variety of similarity metrics. Each similarity value is then used as a feature and trained to predict the human assigned score for that candidate translation. They show that the scores produced by their regression metric using only system-based references correlates with human judgments to the same extent as scores produced using multiple human reference translations. Also, when the regression method was used with human references and some pseudoreferences put together, the correlations obtained by the final metric was better than using the human references alone.

In Albrecht and Hwa (2007), pseudoreferences of different quality—best, moderate and worst—are chosen using the gold-standard judgments and evaluated for use as pseudoreferences. They found that having the best systems as pseudoreferences worked best, although even adding the worst system as pseudoreference gave reasonable performance as their regression approach is trained to predict quality by comparison to the standard of the reference. In their work, however, pseudoreferences of different quality are chosen in an oracle manner (using the human-assigned scores). This setting is not practical because it depends on the actual system scores. In later work, Albrecht and Hwa (2008) use off-the-self machine translation systems as pseudoreferences and show that they can contribute to good results. This later work is a more realistic set-up and here regression is important because we have no guarantees as to the quality of the off-the-shelf systems on the test data.

A similar idea of augmenting machine output to human gold standard was explored in Madnani et al. (2007) in the context of machine translation (MT). For tuning MT systems, often multiple reference translations are required. Madnani et al. augmented reference translations of a sentence with automatically generated paraphrases of the reference. They found in the experiments that such augmentation helped in MT tuning—the number of reference translations needed could be cut in half and compensated with automatic paraphrases.

### 5.2 Choice of Pseudoreference Systems

For this evaluation, the choice of the pseudoreference system is an important step. In this section, we detail some development experiments that we performed to understand how to best choose such pseudoreferences for the summary evaluation task.

We examined a similar regression approach as followed by Albrecht and Hwa (2007). We chose systems of different quality (best, mediocre, worst) based on the

oracle human-assigned scores. The remaining systems were taken as the evaluation set. For each summary in the evaluation data, we computed features to indicate their similarity with the summaries of the chosen pseudoreference systems. Our similarity features were the recall scores from ROUGE overlaps. We computed one feature each for unigram, bigram, trigram, and four-gram ROUGE scores. Each of these four features is computed for each pseudoreference summary.

The scores are used in a linear regression model to predict the summary score. We used a cross-validation approach where the summaries from one of the evaluation systems were used as the test set and the summaries from the remaining systems are used for training the regression model. Then the average predicted score of each system in the evaluation data was computed and compared with their average scores as assigned during manual evaluations.

The experiment was performed using data from four years of DUC conferences, 2001 to 2004. The manual scores in these earlier DUC years were the content coverage scores (described in Section 2.1), which use a single model summary for comparison. Table 7 shows the Spearman correlations between the scores from evaluations only against the pseudoreferences and those from the manual evaluation with the single model. The different settings for choice of pseudoreference systems are also indicated.

These results showed that using the best systems as pseudoreferences provided the best performance across different years. When only worst or only mediocre systems were used, the performance was much worse for predicting system scores. Even when the best systems were augmented with the worst systems as pseudoreferences, the evaluation quality decreased compared with using the best systems only.

Whereas Albrecht and Hwa (2007, 2008) obtained a slight improvement by also using a worst quality pseudoreference in the mix, for summarization it is better to have only the best systems. One reason for this difference could be that for summary evaluation, examples of worse summaries are not very informative. Two good summaries may have considerable variation in the content. When a summary is similar to a best system, therefore, we can say that the candidate summary is also of good quality. On the other hand, when a candidate summary is similar to a worst system summary, it may either be a worse summary or it may be a good summary with different content than the best system's summary. Indeed, when ROUGE was first introduced, it was heavily emphasized that it is a recall measure and that precision-oriented measures do worse. Hence the weights learned for the similarity with the worst system may not be very informative. In summarization, the space of both good summaries and worse summaries for the same input is large. Having more examples of good summaries appears to benefit evaluation more compared with having samples of worst quality.

**Table 7**
Spearman correlations between pseudoreference-based regression scores and manual content scores. The first column lists the type of pseudoreference chosen.

| Pseudoreference | 2001 | 2002 | 2003 task 2 | 2004 task 2 | 2004 task 5 |
|---|---|---|---|---|---|
| 2 best systems | 0.58 | 0.77 | 0.51 | 0.93* | 0.83* |
| 2 worst systems | 0.45 | −0.94* | −0.09 | 0.13 | −0.23 |
| 2 mediocre systems | 0.72 | 0.08 | 0.53 | 0.64 | 0.18 |
| 2 best, 2 worst systems | 0.38 | 0.20 | 0.25 | 0.92 | 0.73 |

*The correlation was significant with p-value $< 0.05$.

Because the best systems turned out to have the maximum potential for acting as pseudoreferences, we wanted a way to identify some best systems without having to rely on the oracle scores, as before. This idea is feasible for our set-up. In our evaluation, we aimed to augment an existing model, so we used the available model to automatically obtain an idea of some of the good systems from the pool. Then we chose some of these top systems as pseudoreferences and combined them with the one available model to form the reference set for final evaluation. Because the reference set has mostly best summaries, we did not use a regression approach based on similarity to the different references. Rather, we considered all of them as models and computed a single ROUGE score comparing a system summary with the pool of model plus pseudomodel summaries.

### 5.3 Experimental Set-up

We now detail our experiments on the TAC 2009 data.

TAC provides four model summaries for each input. We assume that only one is available and choose a model for each input: the first in alphabetical order of identifier names. Based on this model, we compute the RSU4 scores for all systems. We use two methods to choose the pseudomodel systems.

In the first approach, we rank all the systems based on their average scores over the entire test set. The summaries of the top three overall best systems (*global selection*) are added to the set of models for all inputs. Alternatively, we also investigate a different selection method. For *each input*, the top scoring three summaries are added as models for *that* input (*local selection*). In both cases RSU4 was used to identify the best systems according to the single available gold standard.

The final rankings for all systems are produced using the RSU4 comparison based on the expanded set of models (1 human model + 3 pseudomodel summaries). We implemented a jackknifing procedure so that the systems selected to be pseudomodels (and therefore reference systems) could also be compared to other systems. For each input, one of the reference systems (pseudomodels or human model) was removed at a time from the set of models and added to the set of systems. The scores for the systems were then computed by comparison with the three remaining models. The final score for a system summary (not a pseudomodel) is the mean value of the scores with the four different sets of reference summaries created by the jackknifing procedure. For pseudomodel systems, a single score value will be obtained per input resulting from the comparison with the other three models.

### 5.4 Results

The system and input level performance before and after the addition of pseudomodels is shown in Table 8. The performance using four human models is shown in the last line for comparison.

At the macro level, the pseudomodel summaries provide little improvements. Only for the global model is there an increase in correlation, from 0.80 to 0.82.

As expected, however, for the micro level, pseudomodels prove beneficial. Both global and local selection methods improve the number of inputs that receive significant micro-level correlations with pyramid scores. The improvement is close to 10% compared with using only one model summary. Also note that, after the addition of pseudomodels, the percentage of significant correlations is 93%, which is only 2% less compared with the results using four human models (95%).

**Table 8**
Performance before and after the addition of pseudomodel summaries: TAC'09 data (53 systems).

| | Correlations | | | | Pairwise accuracy | | | |
|---|---|---|---|---|---|---|---|---|
| | Macro level | | Micro level | | Macro level | | Micro level | |
| **Evaluation type** | **py** | **resp** | **py** | **resp** | **py** | **resp** | **py** | **resp** |
| RSU4 - 1 model | 0.92 | 0.80 | 84.1 | 79.5 | 88.3 | 80.3 | 66.1 | 50.7 |
| Global | 0.91 | **0.82** | **93.2** | 79.5 | 88.6 | **83.5** | 66.8 | **51.3** |
| Local | 0.92 | 0.79 | **93.2** | 75.0 | **89.6** | 80.8 | **67.4** | **51.3** |
| RSU4 - 4 models | 0.92 | 0.79 | 95.4 | 81.8 | 88.4 | 80.0 | 70.5 | 53.0 |

For responsiveness scores that are model-independent, however, little improvements are seen at both macro and micro levels. The pairwise accuracy at micro level for responsiveness is 1% better after the addition of the pseudomodels.

Comparing the two methods for selecting the best system that can serve as a pseudomodel, the global selection of the system that performed best over the entire available data set appears to be more desirable. It improves the correlations with pyramid scores while keeping the same correlations with responsiveness as with one model. Local selection provides the same performance as global selection for pyramid scores, although it decreases the micro-level evaluation quality for responsiveness.

## 6. Consensus-Based: Evaluation Using Only Collection of System Summaries

From our experiments with pseudomodels, we see that the addition of system summaries to available models proved beneficial and improved the micro-level performance of ROUGE. One question that arises is whether the collection of system summaries together will be useful for evaluation without any human models at all. Again, this idea is related to model-free evaluation. When several systems are available, we investigate if their collective knowledge can help assess summary quality.

Systems use varied methods to select content, and agreement among systems could be indicative of important information. This intuition is similar to that behind the manual pyramid method: Facts mentioned only in one human summary are less important compared to content that is mentioned in multiple human models. For the experiments reported in this section, we rely entirely on the combined knowledge from system summaries as a gold standard.

### 6.1 Related Work

The closest work to this idea of combining system output can be found in the area of information retrieval (IR). Soboroff, Nicholas, and Cahan (2001) proposed a method for evaluating IR systems without relevance judgments. In addition to requiring less human input, the need for automatic evaluation in IR is also motivated by the fact that for systems such as those on the Internet, the documents keep changing and so it is difficult to collect relevance judgments that are stable and meaningful for a long time.

Soboroff, Nicholas, and Cahan (2001) combine the top $n$ results from all the systems and then sample a certain number of documents from this pool. Those documents selected by many systems are more likely to be in the chosen sample and assumed to be most relevant. The systems are then evaluated by considering this chosen set of documents as the gold-standard relevant set.

In our work, we do not attempt to pick out common content explicitly from the summary pool. If we were to follow the same approach as IR, we would be sampling sentences from the summary pool. But in multi-document summarization, sentences from different documents could contain similar content and we do not want to sample one sentence and use it in the gold standard because then systems would be penalized for choosing other similar sentences. In our work, therefore, we break down the sentences and represent the content as a probability distribution over words. A summary is evaluated by comparing its word distribution to that of the pool. We expect that the distribution would implicitly capture the common content.

### 6.2 Evaluation Set-up

For each input, we collect all the summaries produced by automatic systems and calculate the probabilities of words in the combined set. In this way, we obtain a global probability distribution of words selected in system summaries. In this distribution, the content selected by multiple systems will be more prominent, representing the more important information. The word probabilities from each individual summary are then calculated and compared to the overall distribution using JS divergence. If we assume that system summaries are collectively indicative of important content, then good summaries will tend to have properties that are similar to this global distribution, resulting in low divergence values. We compute the correlations of these divergence values with human-assigned summary scores and Table 9 shows the results from this evaluation.

### 6.3 Results

The correlations are on par with those based on multiple human gold standards. At both macro and micro levels, the correlations and pairwise accuracy are similar to those obtained by ROUGE comparison with four human models. The macro-level correlation is 0.93 with pyramid scores, which is very high for a metric that uses no human input at all. Further, the micro-level correlations are also significant for 90% of the inputs. In our pseudomodel experiments, the gains after the addition of system summaries were

**Table 9**
Performance of consensus evaluation approach on TAC'09 data (53 systems). For input level (micro), the percentage of inputs with significant correlations is reported.

| | **Correlations** | | | | **Pairwise accuracy** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Macro level** | | **Micro level** | | **Macro level** | | **Micro level** | |
| **Evaluation type** | **py** | **resp** | **py** | **resp** | **py** | **resp** | **py** | **resp** |
| SysSumm | 0.93 | 0.81 | 90.9 | 86.4 | 88.8 | 80.7 | 65.2 | 52.7 |
| RSU4 - 4 models | 0.92 | 0.79 | 95.4 | 81.8 | 88.4 | 80.0 | 70.5 | 53.0 |

modest (only at micro level). Here we see that a large collection of system summaries by themselves have the information required for evaluation.

From this experiment, we find that consensus among system summaries is indicative of important content. This result suggests that by combining the content selected by multiple systems, one might be able to build a summary that is better than each of them individually. In fact, this idea of system consensus has been utilized in the development of MT systems for quite some time. One approach in MT is rescoring the *n*-best list from an individual system's decoder, and picking the (consensus) translation that is close on average to all translations. Such rescoring is implemented using a minimum Bayes risk technique (Kumar and Byrne 2004; Tromble et al. 2008). The other approach is system combination where the output from multiple systems is combined to produce a new translation. Several techniques including minimum Bayes risk have been applied to perform system combination in machine translation. Shared tasks on system combination have also been organized in recent years to encourage the development of such methods (Callison-Burch et al. 2010, 2011). Such strategies could be a useful direction to explore for summarization as well.

## 7. Discussion

In this article, we have discussed metrics for summary evaluation when human summaries are not present. Our results have shown that these metrics in fact correlate highly with human judgments. But we also need to understand how robust these metrics are and be aware of their limitations. In this section, therefore, we provide a brief discussion of the use of these metrics in different settings.

### 7.1 Including Input–Summary Similarity or Consensus-Based Measures in a Summarization System

Firstly, because input–summary similarity features are computed using the input, they can be useful features to incorporate in a summarization system. The combination of systems to perform evaluation also provides a way to build a better system. The concern would be how the usefulness of these metrics will change if systems were also optimizing for them. To optimize a metric such as JS divergence exactly would be difficult because the JS divergence score cannot be factored or divided among individual sentences, a necessary condition if the problem should be solved using an Integer Linear Program as in McDonald (2007) and Gillick and Favre (2009). Therefore only greedy methods are possible. In fact, KL divergence was greedily optimized in Haghighi and Vanderwende (2009) to obtain a high performance summarizer. Gaming the evaluation should carry little concern, however, as these metrics are proposed with a view to tuning systems.

The metrics we presented are developed for evaluation in a new setting where model summaries are not available and to aid system development and tuning. Further, notice from the micro-level evaluation that a single metric such as JS divergence does not predict content selection performance well for all inputs. System developers should therefore involve other specialized features as well. Regression of similarity metrics is a better predictor at the micro level but optimizing that would involve computation of all metrics. Another point to note here is that these similarity measures and the consensus pool are only indicative of summary content quality. Other key aspects of summary quality, however, involve sentence ordering, proper generation of referring expressions

and grammatical sentences, and maintaining non-redundancy. Systems should therefore be optimizing for a wide variety of factors and thus input–summary similarity and consensus evaluation can be used in the final output to measure the content quality of the summary. Any content evaluation should obviously be accompanied by linguistic quality evaluation in contrast to the current trend to only report content scores.

The high performance of the JS divergence metric also has another implication for system development. On average, the JS divergence measure is highly predictive of summary quality. It indicates that for a large number of inputs in the TAC data sets, good content can be predicted with high accuracy just based on the input's term distributions. Such inputs should therefore be easy to summarize for systems. Although discourse-based and other semantic approaches to summarization have been proposed, most of the systems in TAC rely on surface features such as word distributions. In this situation, we may not be focusing on robust systems that can handle a variety of inputs. In the early years of DUC, the test set comprised a variety of inputs such as biographies, collections of multiple events, opinions, and descriptions of single events. Later years switched to more single-event-type test sets. The results from our analysis point out that current inputs might be too simple for systems and that the range of inputs in the TAC conference should be expanded to include some input types where more sophisticated methods become necessary. Perhaps the input–summary similarity metrics will be helpful in picking out those inputs that need deeper analysis. In the following section, we provide some further analysis into the cases where the input–summary similarity turns out less predictive.

### 7.2 Input–Summary Similarity and Dependence on Input Characteristics

JS divergence is useful for the average rating of systems on the test set and, in our case, we have 44 examples over which the scores are averaged. At the micro level, certain inputs received poor evaluations from JS divergence. Here we provide some insights into the types of inputs where JS divergence worked and the cases which proved difficult.

Table 10 shows the titles of articles in input D0913, the input that received the best evaluation from JSD (correlation of 0.86). These articles were all published on the same day and deal with the same event, a Supreme Court hearing of a case. This input can be said to be highly cohesive and to be discussing the same topic. For such inputs, the term distribution in the input would reflect content importance since some words have higher probability than others because they are discussed repeatedly in the input documents. Such a term distribution when compared with summaries will give good evaluation performance. We can also see that the human summaries for this input (also shown in Table 10) seem to report the common issues observed in the input. In this case, therefore, input–summary similarity scores can predict the pyramid scores that were assigned based on the model summaries. We also show in the table the summary that is chosen to be best according to JS divergence and the summary that had the worst score. We find that the best summary indeed conveys some of the main issues also reported in the human summaries. On the other hand, the low-scoring summary presents a story line about one of the lawyers involved in the case, which is a peripheral topic described in only one of the input documents. In fact, the summary scored as worst by JS divergence has a pyramid score of 0, whereas the chosen best summary has a pyramid score of 0.39.

On the other hand, summaries for input D0940 obtained only 0.3 correlation using JSD evaluation. Both ROUGE and consensus evaluation (SysSumm) methods can

**Table 10**
Titles of articles and two human summaries for input D0913-A. The summaries chosen as best and worst according to JS divergence are also listed.

**Articles in input D0913-A**

| Publication date | Title |
|---|---|
| Mar 02 | US Supreme Court examines issue of displaying Ten Commandments |
| Mar 02 | US Supreme Court examines Ten Commandments displays |
| Mar 02 | Supreme Court wrestles with Ten Commandments issue |
| Mar 02 | Justices examine 10 Commandments case |
| Mar 02 | High Course argues in 10 Commandments case |
| Mar 02 | Supreme Course wrestles with Ten Commandments |
| Mar 02 | High Court argues Ten Commandments cases |
| Mar 02 | Texas seeks to keep ten commandments display on capitol grounds |
| Mar 02 | An unlikely journey up the legal ladder |

**Model summary H**

The Supreme Court heard arguments in two cases on March 2nd about the conditions under which the government could display the Ten Commandments and whether such displays violated the First Amendment.
The Texas case concerns a 40-year-old granite monument of the Ten Commandments, one of 17 monuments on the grounds of the state capital.
The Kansas case concerns framed copies of the Ten Commandments which hang with non-religious documents in two county courthouses.
The displays have been modified twice in response to lower court rulings.
Justice O'Connor is expected to cast the swing votes.

**Model summary C**

On March 2, 2005, the Supreme Court heard two cases concerning the display of the Ten Commandments on government property.
In Texas, a homeless former lawyer challenged the constitutionality of an inscribed monument displayed on state capitol grounds.
In Kentucky, the ACLU claimed copies displayed in two courthouses, modified twice before in response to court rulings, still violated the First Amendment because the original purpose was religious and the modifications were a sham.
Supporters argued that the Ten Commandments are a recognized symbol of law, with both secular and religious functions.
The justices will issue an opinion by late June.

**Lowest JSD summary: System 26**

The Supreme Court Wednesday wrestled with whether the Constitution allows displays of the Ten Commandments on government property.
Lower courts have issued conflicting rulings in dozens of cases involving Ten Commandments displays in recent years.
The justices also heard a challenge to Ten Commandments exhibits in Kentucky courthouses.
Abbott, making his first argument before the high court, contended that the Ten Commandments monument, while containing a sacred religious text.
Abbott, 47, a former state Supreme Court justice.
It is hardly little – a constitutional challenge to displaying the Ten Commandments on the grounds of the Texas Capitol, and so on.

**Highest JSD summary: System 39**

He said his happiest moment was not when he heard the Supreme Court was taking his case but when his teenage daughter read the story and tracked him down by e-mail, breaking a long estrangement.
If religious, it was unacceptable, Van Orden said.
He ate on food stamps at the upscale Central Market, pitched his tent nightly and took it down each morning in a wooded location he did not specify, traveled on a free bus pass granted for a veteran's disability and read newspapers and magazines free at newsstands.

evaluate the same summaries, however, with correlation of 0.84 (ROUGE) and 0.74 (SysSumm). The titles of the articles in that input and in the human summaries are provided in Table 11. This input's topic is the opening of Disneyland in Hong Kong but its articles cover varied aspects around the topic such as ticket sales, environmental

**Table 11**
Titles of articles and two human summaries for input D0940-A. The summaries chosen as best and worst by JS divergence are also listed.

**Articles in input D0940-A**

| Publication date | Title |
|---|---|
| June 22 | Opening of HK Disneyland to be divided into 3 phases |
| June 26 | Disney officials consulted feng shui experts for Hong Kong Disneyland |
| July 01 | Hong Kong Disneyland starts on-line tickets selling |
| July 01 | Disneyland rehearsal days to start in August |
| July 04 | HK Disneyland ticket sale proceeds well |
| Sept 04 | High hopes for Hong Kong Disneyland's economic impact, but critics say Disney magic overrated |
| Sept 08 | Hong Kong Park: Classic Disney with an Asian accent |
| Sept 08 | Hong Kong Disneyland won't cut its maximum capacity despite overcrowding fears |
| Sept 08 | All tickets for opening day of HK Disneyland sold out |
| Sept 10 | Shark fins, stray dogs and smog - Hong Kong Disneyland has had a bumpy ride |

**Model summary B**

Hong Kong Disneyland (HKD) was scheduled to open in three phases in the summer of 2005.
Early to mid-August transportation would be available to some areas of the park.
August-Sept 11 all public services would become available.
Sept 12 the grand opening of the park and hotels would take place.
On Aug 16 HKD will begin its rehearsals to which special guests will be invited.
In September, HKD announced it would not cut its daily capacity of 30,000 visitors.
On Sept 9 it was revealed that all tickets for the grand opening had been sold.

**Model summary H**

Hong Kong Disneyland, a joint venture between Disney and the Hong Kong government, was scheduled to open on 12 September.
Rehearsal days were staged for a month before opening, giving "cast members" a chance to practice their performances.
Disneyland refused to reduce its daily maximum capacity of 30,000 despite complaints from early visitors about large crowds and long lines.
All 16,000 opening day tickets were sold out.
The park is vintage Disney, with aspects of local culture including feng shui, Asian foods, and signs in Chinese.
Protests forced them to remove shark fin soup from their menus.

**Lowest JSD summary: System 54**

But critics say Hong Kong Disneyland is overrated.
The opening of Hong Kong Disneyland is expected to turn on a new page of Hong Kong tourism, with focus on family tourists, she said.
Hong Kong Disneyland will stage its rehearsal from Aug. 16 to the theme park's opening on Sept. 12, the park said in a press release on Friday.
Many in Hong Kong are ready to give Mickey Mouse a big hug for bringing Disneyland to them.
Hong Kong Disneyland Hotel starts at 1,600 (euro 170) a night and Disney's Hollywood Hotel's cheapest room costs 1,000 (euro 106).

**Highest JSD summary: System 1**

Many in Hong Kong are ready to give Mickey Mouse a big hug for bringing Disneyland to them.
But not dog-lovers, shark-defenders and fireworks foes.
The opposition may seem odd, in a Chinese city where fireworks are a fixture, shark fin soup is hugely popular, and stray dogs are summarily dealt with as health hazards.
But eight years after the British colony was returned to China, the capitalist city is much freer than the Communist mainland, and advocacy groups are vocal.

concerns, and use of feng shui. The human summaries for this input focus on different aspects. The term distributions in such an input by themselves do not provide an indication of what was important in contrast to the more cohesive input we discussed previously. The semantics of the content should be better understood to be able to predict the content that humans would choose in their summaries. Subsequently, we can also observe that the summary that is top ranked by JS divergence does not have

much of the same information as the model summaries and talks about yet another set of aspects such as hotel rates and family tourism. The worst summary presents a different set of facts. The pyramid scores for both these summaries are low (0.13 for the best JS summary and 0.0 for the worst JS summary). Input–summary similarity is therefore less helpful here and the information provided by model summaries would be the best gold standard.

Saggion et al. (2010) report that trends can be observed in the JSD metric performance although it does not provide good evaluations for opinion and biographical type inputs. Automatic evaluations in different genres therefore have different requirements and exploring these is an avenue for future work. Input–summary similarity based only on word distribution works well for evaluating summaries of cohesive-type inputs.

We can also envision a situation where we will be able to predict whether the JS divergence evaluation will be accurate or not on a particular test set. In prior work in Nenkova and Louis (2008) and Louis and Nenkova (2009b), we have explored properties of summarization inputs and provided a characterization of inputs into cohesive and less cohesive based on automatic features. The less cohesive inputs were found to be the ones where automatic systems in general performed poorly. In that work, we proposed features to predict if an input is cohesive or not. We now apply these features to the TAC'09 data with the intention of automatically identifying inputs suitable for JS divergence evaluation (the cohesive ones). The features were trained on data from previous years of TAC evaluations. Among the top ten inputs for which JS divergence gave the best correlations, six of them were predicted as cohesive, and, similarly for the bottom ten, six inputs were predicted as "not cohesive." This result provides more validation of the relationship between input and evaluation quality but the automatic prediction of evaluation quality does not appear to be very accurate based on our current features. We plan to explore this direction further in future work.

### 7.3 Requirements for Consensus-Based Evaluation

In a similar vein, one would like to understand the performance guarantees from the consensus-based evaluation method (SysSumm). Here, the metric depends on the availability of a number of diverse system summaries. In the TAC workshops, over 50 systems compete and thus we have a large pool of system summaries with which to compute consensus. For other data sets, when we have to evaluate a few different systems, it is unclear if the same performance can be obtained. To understand the dependence on the number of systems, we study how well the consensus evaluation method works when a small set of standard summarization methods is taken as the available system pool. We expected that when the standard algorithms are chosen to be diverse, their strengths can be combined usefully in a similar manner as the TAC systems.

We choose a set of nine different summarization approaches. They are briefly described here.

**Baseline:** One of the commonly used baseline approaches for multi-document summarization. The first sentence from each document in the input is first included in the summary. After including the first sentence from each document, the second sentence is included and so on up to the length limit.

**Mead:** Radev et al. (2004a, 2004b) rank sentences using a combination of three aspects (sentence length, position in the article, and a centroid score which indicates how central the content of the sentence is) computed by comparison with all other sentences.

**Average probability:** This is a competitive summarizer (Nenkova, Vanderwende, and McKeown 2006) using only the frequency of words as the indicator of content importance. We implement this method by first computing the unigram probability of all content words in the documents of the input combined together. Then we score each sentence by the average value of the probability for the content words in that sentence.

**Topic word:** This is a strong, yet simple, method for generic summarization (i.e., the set of documents given as input must be summarized to reflect the sources as best as possible; in contrast, TAC 2009 tasks can be considered as focused summarization where either a query is provided or an update is required). This method first computes a set of topic words from the input using a loglikelihood ratio. The sentences are ranked using the score introduced by Conroy, Schlesinger, and O'Leary (2006): the ratio of the number of unique topic words in the sentence to the unique content words in the sentence.

**Graph centrality:** This approach (Erkan and Radev 2004; Mihalcea and Tarau 2005) performs selection over a graph representation of the input sentences. Each sentence is represented in vector space using unigram word counts. Two sentences are linked when their vectors have a cosine similarity of at least 0.1. When this graph is converted into a Markov chain, we can compute the stationary distribution of the transition matrix defined by the graph's edges. This stationary distribution gives the probability of visiting each node during repeated random walks through the graph. The high probability nodes are the ones that are most visited and these correspond to central sentences for summarization. The probability from the stationary distribution is the ranking score for this method.

**Latent Semantic Analysis (LSA):** The LSA technique (Deerwester et al. 1990) is based on the idea of dimensionality reduction. For summarization, first, the input is represented as a matrix indexed by words (rows) and sentences (columns) and each cell indicates the count of the word in that sentence. This matrix is converted by singular value decomposition and dimensionality reduction to obtain a matrix of sentences versus concepts where concepts implicitly capture sets of co-occurring terms. The number of concepts is much smaller than the size of the vocabulary. The process also produces the singular values that indicate the importance of concepts. Sentences are selected for each concept in order of concept importance up to the summary length limit. We obtained summaries using the approach detailed in Gong and Liu (2001).

**Greedy-KL:** This method selects sentences by minimizing the KL divergence of the summary's word distribution to that of the input. The idea is similar to findings from our input–summary similarity evaluations. Because the selection of sentences that minimize divergence can only be done by examining all combinations of sentences, Haghighi and Vanderwende (2009) introduce a greedy approach that, at each step, adds the sentence $s_i$ to the existing summary $E$ such that the combination $E \cup s_i$ has the lowest KL among all options for $s_i$.

**CLASSY 04:** This system (Conroy and O'Leary 2001) combines the occurrence of topic words and position of the sentence to predict the score for a sentence. In addition, it employs a hidden Markov model–based approach, so that the probability of a sentence being a summary sentence is dependent on the importance of its adjacent sentences in the document. This system was introduced in the DUC evaluations in 2004 (Conroy et al. 2004). We obtained these summaries (and the subsequent CLASSY 11) from the authors.

**CLASSY 11:** This is a query-focused summarization system used by Conroy et al. (2011) in TAC 2011. It uses features related to topic words and other important keywords identified using a graph-based approach. Rather than greedy selection of top sentences, CLASSY 11 solves an approximate knapsack problem to obtain a more globally optimal summary. Further, the scoring in this method uses bigrams as the basic unit/keyword in contrast to the other methods we have described previously that assume that a sentence is composed of a bag of unigrams.

We generated 100 word summaries from each described system. Except for the CLASSY system, which performs more sophisticated redundancy removal, for the other methods we used the greedy Maximum Marginal Relevance technique (Carbonell and Goldstein 1998) for reducing redundancy. After the sentence rankings were obtained, we added each sentence in order if it was not highly similar (a threshold value on cosine overlap is specified to indicate high similarity) to any of the already added sentences.

Each of the original TAC systems summaries were evaluated as follows. We added the candidate summary to the pool of summaries from these other standard methods. Then we computed the JS divergence between the candidate summary and the combined pool to obtain the score for the candidate. The procedure is the same as the one we followed in Section 6 except that here we assumed that for each TAC system, we only had these standard systems as peers rather than the full set of all TAC systems. The results from this evaluation are shown in Table 12 as *SysSumm-std*$^{9}$. The previous evaluation results, using all TAC systems as consensus, is reproduced in the table as *SysSumm-full*.

We found that even with these few systems, the consensus evaluation is rather strong and produces correlations of 0.91 with pyramid and 0.77 with responsiveness scores. These results provide additional support for the argument that high quality evaluation is feasible even with standard systems as peers and that a small set of such systems appears to be sufficient for forming the consensus.

Because the CLASSY systems are currently some of the top performing systems at TAC, we also evaluated how useful the consensus is if the CLASSY summaries are left out. So we evaluated the TAC systems using only the seven other standard systems (i.e., all except CLASSY04 and CLASSY11) as the peers and the results from this evaluation are reported in Table 12 as *SysSumm-std*$^{7}$. We find that the correlations

**Table 12**
Performance of consensus evaluation approach on TAC'09 data (53 systems). For input level (micro), the percentage of inputs with significant correlations is reported. The results using TAC'09 systems as pseudomodels are indicated as SysSumm-full and those with off-the-shelf systems as SysSumm-std.

| | **Correlations** | | | | **Pairwise accuracy** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Macro level** | | **Micro level** | | **Macro level** | | **Micro level** | |
| **Evaluation type** | **py** | **resp** | **py** | **resp** | **py** | **resp** | **py** | **resp** |
| SysSumm - full | 0.93 | 0.81 | 90.9 | 86.4 | 88.8 | 80.7 | 65.2 | 52.7 |
| SysSumm - std$^{9}$ | 0.91 | 0.77 | 86.3 | 75.0 | 87.4 | 78.7 | 66.4 | 50.9 |
| SysSumm - std$^{7}$ | 0.91 | 0.78 | 84.0 | 75.0 | 87.7 | 78.8 | 65.9 | 50.6 |
| RSU4 - 4 models | 0.92 | 0.79 | 95.4 | 81.8 | 88.4 | 80.0 | 70.5 | 53.0 |

remain the same even when the strongest systems are removed. The usefulness of the consensus therefore is not heavily dependent on the presence of best-quality systems in the pool.

### 7.4 Cross-Year Variation in Metric Performance

The performance of a metric should also be discussed under the effect of different test data sets. In our feature analysis for input–summary similarity, we found that JS divergence produces very high correlations on the TAC 2008 data, about 0.88 with pyramid scores. The performance on the TAC 2009 data, however, although still high (0.74), is lower than the previous year's data. Such a difference could be attributed to different factors. One possible factor is the cohesiveness of the input, as we have discussed earlier. In the TAC evaluations, there is no control over the input types, so inputs from different years may not have the same characteristics. It could be, therefore, that TAC 2009 had more inputs that were less cohesive as our second example above compared to inputs that have more homogenity in the topic discussed. A further evidence for this hypothesis can be seen from the cross-year performance of different metrics presented in Table 13.

Here, improved correlations from 2008 to 2009 are bolded and those that decreased are italicized. The correlations with pyramid scores increased for SysSumm and ROUGE evaluations but dropped for JS divergence. This trend indicates that the model-based evaluations (SysSumm has characteristics similar to model-based evaluation) have more strength on the 2009 data. For inputs where model information cannot be obtained from the general term distribution in the inputs (as could have been the case in 2009), therefore, input–summary similarity that is model-free obtains worse performance. This model dependence can also explain why ROUGE and SysSumm have lower correlations with responsiveness in 2009 despite being able to predict pyramid scores better. Because ROUGE scores are computed based on these specific models, its correlations with the model-free responsiveness judgments drops in 2009.

## 8. Conclusion and Future Work

We have presented successful metrics for summary evaluation that require very little or no human input. We have explored two scenarios: fewer model summaries and no model summaries at all. For both these cases, our newly proposed evaluation metrics have provided good performance.

We analyzed two methods for evaluation in the absence of gold-standard summaries. One was based on input–summary similarity. We examined different

**Table 13**
Cross-year system level correlations for different metrics. The ROUGE-SU4 scores use all four human summaries for reference. Improved correlations in 2009 are **bolded** and decreases in correlations are *italicized*.

| | JSD | | SysSumm | | ROUGE-SU4 | |
|---|---|---|---|---|---|---|
| **year** | **py** | **resp** | **py** | **resp** | **py** | **resp** |
| 2008 | 0.89 | 0.74 | 0.85 | 0.82 | 0.88 | 0.83 |
| 2009 | *0.74* | *0.70* | **0.93** | *0.81* | **0.92** | *0.79* |

possibilities for measuring similarity and quantified their accuracy in predicting human-assigned scores. Our results showed that the strength of features varies considerably. The best metric is JS divergence, which compares the distribution of terms in the input and summary. Combination of JS divergence with other metrics such as cosine similarity and topic word features also gave high correlations with human scores, around 0.77.

Another method we have introduced is the addition of pseudomodel system summaries to the model set when the number of models is low. Our aim here was to improve the micro-level evaluations and our results show that improvements along this line were provided by the pseudomodels.

We also proposed a model-free metric that measures the similarity of a system summary with the collection of all system summaries for that input. This method actually provided even better performance (0.93 correlations with pyramid scores), which is competitive with ROUGE scores computed using four human models.

Furthermore, our evaluations provide consistent performance. In Louis and Nenkova (2009c), we report the correlations for adjacent years showing that our metrics produce reliable performance for two consecutive years of TAC evaluation and for two tasks, query and update summarization.

These evaluation methods highlight considerations that have received little attention so far and give indications of how to perform evaluations on non-standard test sets with little human input. The situation of having only one model summary is not uncommon and so are test sets where there are no model summaries at all. Here, one could use our proposed approaches in system development and then at a later stage use manual evaluations on a small test set to confirm the results. Further, our metrics also provide valuable insights for system development. From our results, it is evident that optimizing for input–summary similarity using an information-theoretic measure such as JS divergence and optimizing for topic signatures are indeed good approaches for building a generic summarization system. In addition, the results from consensus evaluation show that combining summaries from different systems has the potential of creating a system better than the pool. Currently, more than 50 systems compete in the TAC summarization tasks and we want to explore system combination techniques over their summaries in future work.

We also plan to focus on the incorporation of both content and linguistic quality for evaluation. As we already saw, the correlation between system rankings based on pyramid and responsiveness scores is only 0.85. Furthermore, the correlations of ROUGE as well as our metrics are lower with responsiveness compared with the pyramid. Content scores should therefore always be used together with assessments of linguistic quality, and combining both scores would be necessary for obtaining better correlations with responsiveness.

**Acknowledgments**

We would like to thank John Conroy for providing us with the summaries from the CLASSY system and Xi Lin for the implementation of the LSA-based summarizer. We would also like to thank the reviewers for their comments; in particular, the expanded evaluation of the consenus-based metric was added based on their feedback. The work has been partly supported by an NSF CAREER award (09-53445).

**References**

Albrecht, Joshua and Rebecca Hwa. 2007. Regression for sentence-level MT evaluation with pseudo references. In *Proceedings of ACL*, pages 296–303, Prague.

Albrecht, Joshua and Rebecca Hwa. 2008. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation, ACL*, pages 187–190, Columbus, OH.

Best, D. J. and D. E. Roberts. 1975. Algorithm as 89: The upper tail probabilities of Spearman's rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(3):377–379.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar Zaidan. 2011. Findings of the 2011 workshop on statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 22–64, Edinburgh.

Carbonell, Jaime and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR*, pages 335–336, Melbourne.

Conroy, John M., Jade Goldstein, Judith D. Schlesinger, and Dianne P. O'Leary. 2004. Left-brain/right-brain multi-document summarization. In *Proceedings of the 4th Document Understanding Conference (DUC'04)*, Boston, MA. Available at: `http://duc.nist.gov/pubs/2004papers/ida.conroy.ps`.

Conroy, John M. and Dianne P. O'Leary. 2001. Text summarization via hidden Markov models. In *Proceedings of SIGIR*, pages 406–407, New Orleans, LA.

Conroy, John M., Judith D. Schlesinger, Jeff Kubina, Peter A. Rankel, and Dianne P. O'Leary. 2011. Classy 2011 at TAC: Guided and multi-lingual summaries and evaluation metrics. In *Proceedings of TAC*, Gaithersburg, MD. Available at: `http://www.nist.gov/tac/publications/2011/participant.papers/CLASSY.proceedings.pdf`.

Conroy, John M., Judith D. Schlesinger, and Dianne P. O'Leary. 2006. Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING-ACL*, pages 152–159, Sydney.

Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the Americal Society for Information Science*, 41(6):391–407.

Donaway, Robert L., Kevin W. Drummey, and Laura A. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, pages 69–78, Seattle, WA.

Erkan, Güneş and Dragomir R. Radev. 2004. Lexpagerank: Prestige in multi-document text summarization. In *Proceedings of EMNLP*, pages 365–371, Barcelona.

Gillick, Dan and Benoit Favre. 2009. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, CO.

Gillick, Dan and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles, CA.

Gong, Yihong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of SIGIR*, pages 19–25, New Orleans, LA.

Haghighi, Aria and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of HLT-NAACL*, pages 362–370, Boulder, CO.

Harman, Donna and Paul Over. 2004. The effects of human variation in DUC summarization evaluation. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 10–17, Barcelona.

Jing, Hongyan, Regina Barzilay, Kathleen Mckeown, and Michael Elhadad. 1998. Summarization evaluation methods: Experiments and analysis. In *AAAI Symposium on Intelligent Summarization*, pages 60–68, Palo Alto, CA.

Kumar, Shankar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of HLT-NAACL*, pages 169–176, Boston, MA.

Lin, Chin-Yew. 2004a. Looking for a few good metrics: Automatic summarization evaluation-how many samples are enough. In *Proceedings of the NTCIR Workshop*, volume 4, pages 1–10, Tokyo.

Lin, Chin-Yew. 2004b. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL Text Summarization Workshop*, pages 74–81, Barcelona.

Lin, Chin-Yew, Guihong Cao, Jianfeng Gao, and Jian-Yun Nie. 2006. An information-theoretic approach to automatic evaluation of summaries. In *Proceedings of HLT-NAACL*, pages 463–470, New York, NY.

Lin, Chin-Yew and Eduard Hovy. 2000. The automated acquisition of topic signatures for text summarization. In *Proceedings of COLING*, pages 495–501, Saarbrücken.

Lin, Chin-Yew and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*, pages 71–78, Edmonton.

Louis, Annie and Ani Nenkova. 2008. Automatic summary evaluation without human models. In *Proceedings of TAC*, Gaithersburg, MD. Available at: `http://www.nist.gov/tac/publications/2008/additional.papers/Penn.proceedings.pdf`.

Louis, Annie and Ani Nenkova. 2009a. Automatically evaluating content selection in summarization without human models. In *Proceedings of EMNLP*, pages 306–314, Singapore.

Louis, Annie and Ani Nenkova. 2009b. Performance confidence estimation for automatic summarization. In *Proceedings of EACL*, pages 541–548, Athens.

Louis, Annie and Ani Nenkova. 2009c. Predicting summary quality using limited human input. In *Proceedings of TAC*, Gaithersburg, MD. Available at: `http://www.nist.gov/tac/publications/2009/participant.papers/UPenn.proceedings.pdf`.

Madnani, Nitin, Necip Fazil Ayan, Philip Resnik, and Bonnie J. Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127, Prague.

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

McKeown, Kathy, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Barry Schiffman, and Simone Teufel. 2001. Columbia multi-document summarization: Approach and evaluation. In *Proceedings of DUC*, New Orleans, LA. Available at: `http://www-nlpir.nist.gov/projects/duc/pubs/2001papers/columbia_redo.pdf`.

Mihalcea, Rada and Paul Tarau. 2005. Multi-document summarization with iterative graph-based algorithms. In *Proceedings of the First International Conference on Intelligent Analysis Methods and Tools (IA 2005)*, McLean, VA.

Nenkova, Ani and Annie Louis. 2008. Can you summarize this? Identifying correlates of input difficulty for multi-document summarization. In *Proceedings of ACL-HLT*, pages 825–833, Columbus, OH.

Nenkova, Ani and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL*, pages 145–152, Boston, MA.

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

Nenkova, Ani, Lucy Vanderwende, and Kathleen McKeown. 2006. A compositional context sensitive multi-document summarizer: Exploring the factors that influence summarization. In *Proceedings of SIGIR*, pages 573–580, Seattle, WA.

Owkzarzak, Karolina and Hoa Trang Dang. 2009. Evaluation of automatic summaries: Metrics under varying data conditions. In *Proceedings of the Workshop on Language Generation and Summarisation*, pages 23–30, Singapore.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Radev, Dragomir, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004a. MEAD—A platform for multidocument multilingual text summarization. In *Proceedings of LREC 2004*, pages 1–4, Lisbon.

Radev, Dragomir, Hongyan Jing, Malgorzata Sty, and Daniel Tam. 2004b. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938.

Radev, Dragomir and Daniel Tam. 2003. Single-document and multi-document

summary evaluation via relative utility. In *Proceedings of CIKM*, pages 508–511, New Orleans, LA.
Rath, G. J., 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.
Saggion, Horacio, Juan-Manuel Torres Moreno, Iria da Cunha, Eric SanJuan, and Patricia Velazquez-Morales. 2010. Multilingual summarization evaluation without human models. In *Proceedings of COLING*, pages 1,059–1,067, Beijing.
Soboroff, Ian, Charles Nicholas, and Patrick Cahan. 2001. Ranking retrieval systems without relevance judgments. In *Proceedings of SIGIR*, pages 66–73, New Orleans, LA.
Tromble, Roy W., Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of EMNLP*, pages 620–629, Honolulu, HI.
van Halteren, Hans and Simone Teufel. 2003. Examining the consensus between human summaries: Initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL DUC on Text Summarization Workshop*, pages 57–64, Edmonton.
Zhang, Ying and Stephan Vogel. 2010. Significance tests of automatic machine translation evaluation metrics. *Machine Translation*, 24(1):51–65.