Predicting Summary Quality using Limited Human Input

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

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

Abstract

We present four experiments with summary evaluation approaches that use little or no human input in the form of model summaries or human judgements. We investigate whether *system-produced summaries* can be used to improve predictions of summary quality when few or no human summaries are available. We also validate our previous findings that measures of *input-summary similarity* and *input cohesiveness* are predictive of summary quality. We analyze the performance of our methods in predicting the human assigned scores for summarization systems from the 2008 and 2009 Text Analysis Conferences. Input-summary similarity metrics obtain correlations of about 0.7 with manual pyramid scores on the TAC '09 data. Using only a collection of system summaries in place of gold standard, the correlation is 0.9. We also show that properties of input cohesiveness can predict the average system score with good accuracies.

1 Introduction

Current evaluation methods such as manual pyramid scores (Nenkova et al., 2007) or automatic ROUGE metric (Lin and Hovy, 2003) use multiple human summaries as reference. It is desirable that evaluation of similar quality be done quickly and cheaply on non-standard test sets that have few or no human summaries.

In our work, we aim to identify measurable indicators of summary content quality. We present four experiments using resource-poor approaches to predict summary content scores assigned by human judges. Our methods address the following questions:

What *input-summary similarity* metrics are most useful and consistently predictive of content quality?

Can summaries of a few best systems be used as *pseudomodels* and expand the set of available model summaries on non-standard test sets?

Can a *collection of large number of system summaries* be used as gold standard for evaluation?

Are *input difficulty* features predictive of expected system performance in the query and update summarization tasks? Multi-faceted inputs which have documents with less redundancy and covering diverse topics have been found to be difficult for generic summarization systems in past evaluations.

We evaluate our predictions on data from the Text Analysis Conferences (TAC)[1] conducted in '08 and '09. Two input-summary similarity metrics[‡] and two pseudomodel metrics were submitted for the official TAC '09 task on developing automatic evaluation metrics–AESOP (Automatically Evaluating Summaries of Peers).

We find that our automatic methods to estimate summary quality are highly predictive of human judgements. Our best result is 0.93 correlation with human rankings using no model summaries at all.

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

[‡] Our tool to obtain input-summary similarity scores using different metrics is available at http://www.cis.upenn.edu/ lannie/IEval.html

Our methods provide some direction towards alternative methods of evaluation on non-standard test sets with little human input. Our results also provide insights into how system features to perform content selection can be improved.

2 Data

For our experiments, we use system evaluations from the Update Summarization tracks in TAC 2008 and 2009[2].

The test set for this task is comprised of 48 inputs in '08 and 44 in '09. Every input consists of two sets of documents–docsets A and B. In addition, a query is also provided which expresses the user's information need. For the *query* task, a system must produce a query-focused summary of docset A. In the second *update* task, the system must assume that the user has read all the documents in A and produce a summary of updates for the user from docset B. The target length of the summary is 100 words for both tasks.

2.1 Systems

There were 58 and 53 *automatic* systems that participated in the evaluations in '08 and '09 respectively. Our methods use the input and system summary texts for evaluation, so they tend to provide better scores for extractive system summaries and do not work well for human-written abstractive summaries. We also investigate the possibility of using the output of automatic systems as (part of) the gold standard for evaluation. Because of these two reasons, we limit our analysis to only the automatic systems in these years. The automatic baselines are included in our analysis. Two oracle baselines were used in TAC '09 which were human-produced summaries. We exclude these from our experiments.

2.2 NIST evaluation

The summaries produced by systems were evaluated using two manual methods, pyramid and responsiveness, and an automatic method, ROUGE.

Pyramid scores: The pyramid method (Nenkova and Passonneau, 2004; Nenkova et al., 2007) uses multiple human models for evaluation. It is based upon the intuition that the information expressed in multiple human summaries can be considered more important and central for evaluation purposes compared to content that is mentioned only in one of many human summaries or not mentioned in any. Annotators identify Summary Content Units (SCU) expressed in the human models. Each SCU is assigned a weight equal to the number of model summaries in which it is expressed. An ideal summary would express a subset of the most highly weighted SCUs, with multiple maximally informative summaries being possible.

Four human summaries were used in the pyramid evaluations in TAC '08 and '09. The pyramid score for a system summary is equal to the ratio between the sum of weights of SCUs expressed in a summary (again identified manually) and the sum of weights of an ideally informative summary with the same number of SCUs.

Responsiveness: Responsiveness evaluation does not use human models and is a collective score of both content selection and linguistic quality of a summary. Human judges directly provide ratings of summary quality on a given scale (1 - 5 in TAC '08 and 1 - 10 in '09). The judgements are based upon how well the summary satisfies the information need of the user.

ROUGE: The standard automatic evaluation metric for summarization is ROUGE (Lin and Hovy, 2003; Lin, 2004). It compares a system summary with model summaries automatically using n-gram overlap statistics. Similarity scores on the basis of these overlaps have been shown to correlate well various manual evaluation metrics used in DUC/TAC.

2.3 AESOP '09

In TAC 2009, the AESOP (Automatically Evaluating Summaries of Peers) track was introduced. The goal of the track was to identify automatic metrics for summary evaluation which correlate well with human judgments. The test set for this track consisted of the source documents, topic statements, system output and model summaries from the Update Summarization task for the same year. Participants could use these resources to produce a score for each summary using their automatic method. The scores produced by these automatic metrics

[2] We also use evaluations from DUC 2002 to 2004 as training examples for our input difficulty experiments.

were then checked for correlations with the pyramid and responsiveness scores given by human assessors. All submitted systems except ours sought to improve ROUGE by refining methods for comparing a peer summary with several gold standard summaries produced by humans.

ROUGE-SU4 and BE (Basic Elements) (Hovy et al., 2005) metrics were the official baselines for AESOP ’09. In our work, we use ROUGE-SU4 (RSU4) scores to show how much our methods’ performance differs from the model-based evaluation metric. Our goal is to produce stable evaluation in contexts where few or no model summaries are available to do model-based evaluation.

RSU4 uses skip bigram overlaps to measure similarity between a model and a system summary. Skip bigrams allow for some flexibility in matching whereby non-consecutive words in sentence order within a specified gap limit can be considered as bigrams. RSU4 uses a gap of 4 words at maximum. The scores were computed after stemming all content words. Stop words were retained in the summaries[3].

3 Evaluating predictions of summary quality

We report the performance of our methods in replicating human-assigned overall rankings of systems on the test set as well as their capacity to identify good and bad summaries for individual inputs.

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.

Input level (micro): For each individual input we compare the rankings for the different system summaries obtained by manual and automatic evaluations. Since the correlations are computed for each input, we report the percentage of inputs for which significant correlations were obtained. This analysis highlights the ability of an evaluation to identify good and poor quality system summaries produced for a specific input.

In the following sections, we describe four experiments in which we analyze the possibility of performing automatic evaluation involving only minimal or no human judgements–using input-summary similarity (Section 4), using system summaries as additional pseudomodels with human references (Section 5), using a meta-model composed of system summaries only (Section 6) and using input difficulty features to predict average system scores (Section 7). In evaluating all these methods, the oracle baselines from TAC ’09 were excluded. But all the automatic systems including baselines were evaluated. Two input-summary similarity metrics and two metrics based on pseudomodels were submitted to the official AESOP ’09 track.

4 Input-Summary similarity

In TAC ’08 we presented a model-free evaluation method based upon input-summary similarity. Since summaries are expected to be surrogates of the input, input-summary similarity is an intuitive measure of summary quality (Donaway et al., 2000). But there are multiple ways to measure similarity (Oliveira et al., 2008; Haghighi and Vanderwende, 2009).

We analyzed several similarity features to compare a summary with its input, described in Louis and Nenkova (2008) and Louis and Nenkova (2009a). The similarity scores from our top features obtained very high correlations with human judgements on the TAC ’08 data.

We briefly describe these metrics below and analyze their predictiveness on both TAC ’08 and ’09 data.

4.1 Similarity metrics

Information-theoretic: These include Kullback-Leibler divergence and Jensen-Shannon (JS) divergence between vocabulary distributions of the input and summary. Since good quality summaries are more likely to closely follow the word distributions in the input, we can expect these summaries to have lower divergence compared to poor quality summaries.

[3] We report ROUGE-SU4 scores obtained using ROUGE-1.5 with the following parameters
-n 4 -w 1.2 -m -2 4 -u -c 95 -r 1000 -f A -p 0.5 -t 0 -a -d.

		Macro level				Micro level			
		TAC 2008		TAC 2009		TAC 2008		TAC 2009	
Task	**Evaluation**	**py**	**resp**	**py**	**resp**	**py**	**resp**	**py**	**resp**
Query	JSD	0.89	0.74	0.74	0.71	77.08	72.92	84.09	75.00
	Regr	0.86	0.68	0.77	0.67	77.08	72.92	81.82	65.91
Update	JSD	0.84	0.77	0.72	0.61	85.42	75.00	77.27	72.73
	Regr	0.80	0.73	0.71	0.54	81.25	58.33	75.00	52.27

AESOP ’09 Run1 - JSD, Run2 - Regr

Table 1: Input-summary similarity: macro level–Spearman correlations, micro level–percentage of inputs with significant correlations

Vector space similarity: Cosine similarity is frequently used to compare the content of two texts. We used tf*idf representations of input and summary content words in the two vectors for comparison. Good summaries are likely to have higher similarity values.

Generative model based on frequency: One way to view summary production is 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 assumption, the likelihood of a summary’s content can be computed using different methods and would be higher for better quality summaries. We experimented with unigram and multinomial summary probabilities.

Use of topic signatures: All the above described metrics use the full set of input and summary content words for comparison. In contrast, we also experimented with a restricted set of words from the input–the topic words (Lin and Hovy, 2000). Three features were used, a) the percentage of summary content words which match the input’s topic words, b) the percentage of input’s topic words that also appear in the summary and c) cosine overlap between input’s topic words and summary’s content words.

Regression-based combination: A combination of all the above features using linear regression.

4.2 Results

On the TAC 2008 data, JS divergence and regression metric obtained the best correlations with human rankings for both types of manual scores and summarization tasks. JS divergence obtained correlations in the range of 0.74 to 0.89 and regression around 0.68 to 0.86 across both query and update tasks. These results are shown in Table 1.

We evaluated the systems from the TAC ’09 query and update task[4] using these two top features. For regression, we used cross validation for the ’08 data; for evaluation of TAC ’09 systems, we used the full data from TAC ’08 for training the regression model.

The results (Table 1) validate our prior findings. At system level, the regression metric obtained correlations of 0.77 and 0.71 with pyramid scores for the TAC ’09 query and update tasks respectively. The correlations between pyramid metric and JS divergence scores are 0.74 and 0.72 for query and update summaries.

These results show that our top features obtain fairly consistent performance across the two years. Recent work (Oliveira et al., 2008) also shows that input-summary similarity computed based on word frequency and n-gram overlap information is highly predictive of human rankings of single document summaries from earlier years of DUC. Hence input summary similarity can be used for evaluation with good results when model summaries are not available.

5 Use of pseudomodel system summaries

Methods such as pyramid employ multiple human summaries to avoid bias in evaluation when using a single model as gold standard. ROUGE metrics are

[4] In previous analyses we found that JS divergence with only the update input provided the best predictions for the update summarization task. For regression, the feature values from both background and update inputs were used.

		Macro level				Micro level			
		TAC 2008		TAC 2009		TAC 2008		TAC 2009	
Task	**Evaluation**	**py**	**resp**	**py**	**resp**	**py**	**resp**	**py**	**resp**
Query	RSU4 - 1 model	0.78	0.75	0.92	0.80	79.17	72.92	84.09	79.54
	RSU4 - 4 models	**0.88**	**0.83**	0.92	0.79	**100**	**93.75**	**95.45**	81.82
Update	RSU4 - 1 model	0.91	0.88	0.80	0.69	87.50	79.17	86.36	75.00
	RSU4 - 4 models	0.93	0.90	**0.85**	0.69	**100**	**93.75**	**100**	**86.36**

Table 2: ROUGE evaluation with different number of models: macro level–Spearman correlations, micro level–percentage of inputs with significant correlations

also currently used with multiple summaries, when available. But often few model summaries if any, are available on non-standard test sets. We explored the possibility of predicting the best systems with the few available models and then using summaries of these systems as additional "pseudo-models". Evaluation using system output has been shown to correlate with human scores for machine translation (Albrecht and Hwa, 2007; Albrecht and Hwa, 2008).

5.1 Effect of number of models on evaluation quality

Previous studies (Harman and Over, 2004; Owkzarzak and Dang, 2009) have shown that at the system level, rankings even with a single model will be stable when computed over a sufficient number of test inputs. However, multiple models are particularly important for evaluation at the level of individual inputs.

This effect can be observed from the difference in ROUGE correlations with manual scores when using only one or using four models (Table 2). When a single model summary was used, we choose the first model in alphabetical order of their names.

At system level, the correlations from both setups are often similar. But at micro level, there is considerable difference in performance. Using all four models, significant correlations are obtained for pyramid scores for nearly all inputs in the two years. However, the evaluations using a single model produce significant correlations for only 79 to 87% of the inputs.

5.2 Selection of pseudomodel systems

One cheap method to add additional summaries to the model pool would be to include those produced by a few good systems. We used the one model available per input (chosen as described earlier) to evaluate the systems and obtain rankings. We then chose as pseudomodels the top ranking three systems on the basis of average scores over the entire test set. RSU4 was used to compute similarity between the system and model summaries. The summaries of these overall best systems (*global selection*) were added to the set of models for all inputs. Alternatively, for each input, the top scoring three summaries for that input were added as models for that input (*local selection*).

5.3 Evaluation using expanded model set

The final rankings for all systems were produced using 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 peers. The scores for the peers were then computed by comparison with the three remaining models. The final score for a peer summary (not a pseudomodel) is the mean value of the scores with the four different sets of reference summaries. 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 results from using a single human model and that after the addition of global and locally selected pseudomodels are shown in Table 3.

		Macro level				Micro level			
		TAC 2008		TAC 2009		TAC 2008		TAC 2009	
Task	**Evaluation**	**py**	**resp**	**py**	**resp**	**py**	**resp**	**py**	**resp**
Query	RSU4 - 1 model	0.78	0.75	0.92	0.80	79.17	72.92	84.09	79.54
	Global	**0.81**	**0.76**	0.92	**0.86**	70.83	60.42	**86.36**	79.55
	Local	**0.79**	**0.78**	**0.93**	**0.81**	75.00	72.92	**88.63**	77.27
Update	RSU4 - 1 model	0.91	0.88	0.80	0.69	87.50	79.17	86.36	75.00
	Global	0.90	**0.90**	**0.94**	**0.83**	**89.58**	**81.25**	**88.63**	**81.82**
	Local	0.89	**0.89**	**0.89**	**0.80**	**89.58**	75.00	79.55	**77.27**

AESOP '09 Run3 - Global, Run4 - Local

Table 3: Pseudomodel evaluation results: macro level–Spearman correlations, micro level–percentage of inputs with significant correlations

At the macro level, the addition of both global and local pseudomodels leads to some improvements in correlations with manual scores. The best results are seen for the TAC '09 update task where the correlations improve from 0.80 to about 0.9 (with pyramid) and from 0.69 to about 0.8 (with responsiveness). But the improvements for the other cases are small.

The best results from pseudomodels at the input level is seen for TAC '09 update task with respect to responsiveness–a 6% improvement in number of inputs with significant correlations. Around 2 to 4% improvement can be observed in several other setups. But the use of pseudomodels fails to obtain consistent improvements across different tasks or pseudomodel selection methods for micro level evaluation. Again at the macro level, the addition of pseudomodels led to sizable improvements in correlations for the TAC '09 update task. In other cases, there were small improvments if any. Also notice that in the case of TAC '08 query-focused summaries, the addition of globally selected pseudomodels reduces the micro-level performance by about 10%.

But we have seen that a single model will not be consistently indicative of the summary rankings for individual inputs. Assessing the performance of a system in such settings needs further analysis.

6 Evaluation using collective information from system summaries

In our experiments with pseudomodels, we did not obtain consistent improvements from the addition of the best system summaries to the set of model summaries. One question that arises is whether the collection of system summaries together will be useful for evaluation in a "wisdom of the crowds" fashion.

Systems use varied methods to select content and agreement among systems could be highly indicative of important information. The intuition is similar to that behind the manual pyramid method: facts mentioned only in one human summary are less important compared to content that overlaps in multiple human models. Now we rely entirely on the combined knowledge from system summaries.

6.1 Evaluation setup

For each input, the vocabularies of all system summaries were combined to obtain a global probability distribution of words selected in system summaries. Content selected by multiple systems will be more frequent representing the more important information. Each individual summary was then compared to this overall distribution using Jensen-Shannon divergence. If system summaries are collectively indicative of important content, good summaries will tend to have properties which are similar to this global distribution resulting in low divergence values.

6.2 Results

The results are shown in Table 4. At macro level, the correlations with human judgements range between 0.79 and 0.93 across the two tasks in TAC '08 and '09. These values are in fact only slightly lower than the macro level correlations obtained by

		Macro level				Micro level			
		TAC 2008		TAC 2009		TAC 2008		TAC 2009	
Task	**Evaluation**	**py**	**resp**	**py**	**resp**	**py**	**resp**	**py**	**resp**
Query	SysSumm	0.85	0.82	0.93	0.81	81.25	85.42	90.91	86.36
Update		0.91	0.88	0.89	0.79	89.58	83.33	88.64	77.27

Table 4: Evaluation using system summaries only: macro level–Spearman correlations, micro level–percentage of inputs with significant correlations

ROUGE evaluation with four model summaries (Table 2). At micro level, the percentage of inputs with significant correlations using this method is around 77 to 90% across the two years.

The very high correlations with manual judgements obtained from this evaluation experiment suggest that system combination could be a useful direction to explore for summarization. Sentences selected by multiple methods can be included as summary sentences with high confidence. System combination has been widely explored to improve the quality of machine translation output.[5]

7 Input difficulty: predictor of average system score

In prior work (Nenkova and Louis, 2008; Louis and Nenkova, 2009b), we found that some inputs are more difficult for systems to summarize and are characterized by low *average system performance*. Such variability arises because most of the current systems ignore the properties of specific inputs and use a common method to summarize all inputs.

We identified inputs with less redundancy, high vocabulary sizes and low relatedness between documents as more difficult for current systems. Such features were able to identify difficult inputs (average system score is below the mean value) with accuracies above baseline for the generic summarization tasks in DUC 2002 to 2004. These predictions of the average system performance could be made solely on the basis of properties of the input.

We now examine the performance of these features on the TAC '08 and '09 data.

7.1 Features

We divide the inputs into 2 classes - *"easy"* where the average system performance is above the mean value and *"difficult"* for inputs with below mean value average performance. Equal number of inputs are used in both classes.

Six features were identified as significant predictors of difficult inputs (Nenkova and Louis, 2008).

- large vocabulary size
- high entropy vocabulary
- low KL divergence between input and a large random document collection
- low values for average pairwise cosine similarity betweeen documents. All content words were used in the comparison.
- small % of vocabulary consisting of topic words (Lin and Hovy, 2000)
- low values for average pairwise cosine similarity between documents using topic words only

7.2 Results

We trained a logistic regression classifier using the above mentioned features on the 196 multidocument inputs from DUC 2002 to 2004[6]. The task is binary classification of inputs into *easy* and *difficult* classes. Table 5 shows the performance of these features on TAC '08 and '09 data.

Since we use an equal number of inputs in both classes, the random baseline performance is 50% accuracy. The accuracies from input difficulty features are 10% better than the random baseline.

[5] The 2009 EACL Workshop on Statistical Machine Translation included a shared task on system combination.
http://www.statmt.org/wmt09/system-combination-task.html

[6] Note that the evaluation method used in these years is different, scores were estimated using a single model summary. In addition, the task was generic summarization. However this data provided the best match for our test set in terms of number of documents in an input set and the target summary size.

		TAC 2008			TAC 2009		
Task	**Data**	**acc**	**prec**	**recall**	**acc**	**prec**	**recall**
Query	all	64.58	68.18	60.00	61.36	60.86	63.63
	extremes	75.00	85.71	60.00	60.00	62.50	50.00
Update	all	60.42	63.63	56.00	59.09	56.00	66.67
	extremes	70.00	66.67	80.00	75.00	69.23	90.00

Table 5: Classifying easy vs. difficult inputs–overall accuracy, and precision and recall for difficult inputs

Inputs on which the average system performance was close to the mean are likely to be neither easy or difficult. Therefore we also evaluated the accuracy of our predictions on the 10 inputs each with extremely high and low values for average system score. On these inputs (extremes), accuracies increase to 75%.

These results confirm that most systems do have difficulty summarizing multi-faceted inputs. Hence properties of such inputs turn out predictive of average system performance. Specialized content selection methods would be necessary to smooth out the variable system performance on different inputs.

8 Conclusion

We have presented an analysis of automatic predictions of summary quality. Our findings were evaluated and validated using data from two large scale summarization system evaluations organized by NIST in ’08 and ’09. Our resource-poor methods are able to estimate system quality with different degrees of performance.

Using system summaries as pseudomodels in addition to a single model summary did not provide consistent improvements compared to using only the model summary alone. However, we found that a large number of summaries from different systems can be used to collectively distinguish most important content from others and thereby evaluate a given summary for that input. We were also able to validate findings from our prior work. Input-summary similarity measured by information-theoretic features is predictive of summary quality. Also multi-document inputs with less redundancy and content overlap remain difficult for systems in recent years as well.

References

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

Joshua Albrecht 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.

Shared Task: Automatic System Combination. 2009. http://www.statmt.org/wmt09/system-combination-task.html. *Fourth Workshop on Statistical Machine Translation, EACL*.

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

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

Donna Harman 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.

Eduard Hovy, Chin-Yew Lin, and Liang Zhou. 2005. Evaluating DUC 2005 using basic elements. In *Proceedings of DUC-2005*.

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

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

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

Annie Louis and Ani Nenkova. 2008. Automatic summary evaluation without human models. In *Proceedings of the Text Analysis Conference*.

Annie Louis and Ani Nenkova. 2009a. Automatically evaluating content selection in summarization without human models. In *Proceedings of EMNLP*.

Annie Louis and Ani Nenkova. 2009b. Performance confidence estimation for automatic summarization. In *Proceedings of EACL*.

Ani Nenkova and Annie Louis. 2008. Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *Proceedings of ACL-HLT*.

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

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

Paulo C F de Oliveira, Edson Wilson Torrens, Alexandre Cidral, Sidney Schossland, and Evandro Bittencourt. 2008. Evaluating summaries automatically - a system proposal. In *Proceedings of LREC*.

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