Summary Evaluation: Together We Stand NPowER-ed

George Giannakopoulos and Vangelis Karkaletsis

NCSR Demokritos
Institute of Informatics and Telecommunications
GR-15310, Aghia Paraskevi, Attiki, Greece
{ggianna,vangelis}@iit.demokritos.gr

Abstract. Summary evaluation has been a distinct domain of research for several years. Human summary evaluation appears to be a high-level cognitive process and, thus, difficult to reproduce. Even though several automatic evaluation methods correlate well to human evaluations over systems, we fail to get equivalent results when judging individual summaries. In this work, we propose the NPowER evaluation method based on machine learning and a set of methods from the family of "n-gram graph"-based summary evaluation methods. First, we show that the combined, optimized use of the evaluation methods outperforms the individual ones. Second, we compare the proposed method to a combination of ROUGE metrics. Third, we study and discuss what can make future evaluation measures better, based on the results of feature selection. We show that we can easily provide per summary evaluations that are far superior to existing performance of evaluation systems and face different measures under a unified view.

1 Introduction

Summarization research becomes a necessity in the overwhelming amount of information of our age. The effort to achieve good summaries through automated Natural Language Processing (NLP) can be significantly boosted if one can automatically determine whether a generated summary is good or not. Based on this need, summarization system evaluation research has progressed as a new domain of focus for researchers.

For several years the evaluation community has relied on evaluation measures born in or derived from related NLP tasks (e.g., ROUGE [1] and the related BLUE measures [2]). However, several studies, as well as the experience on new summarization tasks, have shown the need for better evaluation measures [3, 4, 5]. This requirement for new measures is related to a variety of needs, ranging from better discrimination between acceptable and good (human-performance) systems [6] to multi-lingual summarization evaluation [7]. Furthermore, even though existing methods of automatic evaluation do well when judging whole systems, they perform average when judging individual summaries [8].

In this work, we try to "stand upon the shoulders" of existing metrics, which have been proposed over the years. We study whether it makes sense to combine

A. Gelbukh (Ed.): CICLing 2013, Part II, LNCS 7817, pp. 436–450, 2013.

existing, language-agnostic evaluation measures into a single, combined evaluation via optimization. If such a combination is effective, we examine if there exists a subset of features that are adequate for the task at hand. We also perform experiments trying to emulate different aspects of summary evaluation (responsiveness, Pyramid score) under the same, unified perspective.

The paper is structured as follows. We present an overview of summary evaluation literature (Section 2). We desribe the NPowER evaluation method (Section 3) and perform various analyses to determine good strategies for evaluation methods combinations (Section 4). We then conclude, summarizing the findings of this work (Section 5).

2 Summary Evaluation Overview

Summary evaluation allows us to identify errors and reiterate or reformulate certain aspects of the process to optimality. While this is common ground, the notion of automatic evaluation is not. For some time now, the domain of automatic evaluation of summaries was only superficially addressed, because many of the required summary qualities could not be automatically measured. Therefore, human judges have been widely used to evaluate or cross-check the summarization processes [9, 10, 3]. Below, we overview different evaluation types and methods.

An evaluation process can be either intrinsic or extrinsic (e.g., [9, 11]). Intrinsic evaluation operates on the characteristics of the summary itself, trying for example to capture how many of the ideas expressed in the original sources appear in the output. On the other hand, extrinsic evaluation decides upon the quality of a summary depending of the effectiveness of using the summary in a specific task. An extrinsic evaluation case is when we use summaries, instead of source texts, to answer a query. The evaluation is then based on whether the answer is equivalent to the answer derived from source texts. On the contrary, using a *gold standard* summary, i.e., a human-generated summary viewed as the perfect output, and estimating the similarity of the summary to the gold standard, is an intrinsic evaluation case (e.g., [12]).

Sparck-Jones argues [13] that the classification of evaluation methods as intrinsic and extrinsic is not enough and proposes an alternative schema of evaluation methods' classification. This schema is based on the degree to which the evaluation method measures performance, according to the intended purpose of the summary. Therefore, defining new classes that elaborate on the definitions of extrinsic and intrinsic, Sparck Jones classifies evaluation methodologies as: semi-purpose, e.g., inspection of proper English; quasi-purpose, based on comparison with models, e.g., n-gram or information nuggets; pseudo-purpose, based on the simulation of task contexts, e.g., action scenarios; full-purpose, based on summary operation in actual context, e.g., report writing.

In [14] we find a comment (part 3.4) referring to intrinsic evaluation, where the authors suggest that 'only humans can reliably assess the readability and coherence of texts'. This statement indicates the difficulty of that kind of evaluation. But do humans perform perfect in the evaluation of summaries? And what does *perfect* account for?

Humans tend to be able to identify good texts, in a qualitative manner. There is an issue of how to make human assessors grade the quality of a text in uniform and objective ways (see for instance [11, 12] for indications of the problem). At this point numerous efforts have pointed out the inter-judge agreement problem [15, 16, 17, 18]. People tend to have similar, but surely not too similar opinions. This led to looking for subjective measures correlated to human subjectivity. In other words, if our measures behave similarly to human evaluation, we will have reached an adequate level of acceptance for our (automatic) quality measures. In [12] partial inter-judge agreement is illustrated among humans, but it is also supported that, despite the above, human judgements generally tend to bring similar results. Thus, perfection is subjective in the summarization domain: we can only identify good enough summaries for a significant percentage of human assessors.

Pyramid evaluation [15] uses humans to evaluate summaries in a controlled process. The humans are called to identify the segments of the original text, from which pieces of the judged summary are semantically derived. In other words, the method makes use of a supposed (and argued) mapping between summary sentences and source documents, where summarization content units (SCUs) are identified. SCUs are minimal units of informative ability that also appear in the summary output. According to the number of human judges agreeing on the origin of an SCU (i.e., the text span that corresponds to the SCU), the SCUs are assigned weights, corresponding to pyramid layers. Thus, the SCUs higher in the pyramid are supposed to be the most salient pieces of information in the original sources. A summary is then evaluated by locating the SCUs present in the summary output and using a summing function to account for the weights. Doing so, two measures are defined: the pyramid score, which corresponds to precision, and the modified pyramid score, which corresponds to recall. Nenkova argues that the above evaluation process can suppress human disagreement and render useful results. Pyramid evaluation was also applied in DUC and TAC, and the use of a new set of directives for evaluators in DUC 2006 provided better results than DUC 2005 [19], though not reaching the effectiveness of automatic methods. This indicates that manual evaluation methods can be highly dependent on the instructions given to the evaluators.

A number of different intermediate representations of summaries' information have been introduced in existing summarization evaluation literature, ranging from automatically extracted s nippets to human-decided sub-sentential portions of text. These representations form the basis for the comparison between summaries. More specifically, the "family" of BE/ROUGE[1] [21, 1] evaluation frameworks, uses statistical measures of similarity, based on n-grams (of words), although it supports different kinds of analysis, ranging from n-gram to semantic [21]. The intuition behind the BE/ROUGE family is that, in order two texts to have similar meaning, they must also share similar words or phrases. One can take into account simple unigrams (single words) in the similarity comparison, or may require larger sets of words to be shared between compared texts. *Basic Elements* (BE) are

[1] See also [20] for the BLEU method on machine translation.

considered to be 'the head of a major syntactic constituent' and its relation to a single dependent. BEs are decided upon in many ways, including syntactic parsing and the use of cutting rules [21]. BEs can be matched by simple string matching, or by semantic generalization and matching, according to the proposed framework [3, 22]. A more recent work [23] uses variations on dependecies and external information (e.g., WordNet) to overcome the problems that arise from different formulations of model summaries.

An alternative to the aforementioned representations is that of the n-gram graphs [24], where mostly n-grams of characters are used to represent documents. Given a set of "gold standard" texts and their n-gram graphs, the similarity (Value Similarity [24]) between the graph of a judged summary ("peer" summary) and the "gold standard" graphs is used as a grade. This approach has offered two main variations:

- the AutoSummENG [24] original approach. This approach calculates the average of the similarities between the peer summary graph and the gold standard graphs. This average is the grade assigned as a score to the peer summary.
- the MeMoG (Merged Model Graph) variation [25]. In this case, the gold standard graphs are merged into a representative graph. Then, the score assigned to the peer is the similarity between the peer summary graph and the representative graph.

Other variations, based on the notion of the n-gram graph, are the Hierarchical Proximity Graphs (HPG) [25] (using a hierarchy of recursive n-gram graphs) and context chains [26] (n-gram graphs based on co-reference chains).

The most recently faced problems of automatic evaluation relate to:

- the ability of evaluation measures to take into account redundancy over subsequent summaries on the same topic. This task ("update" task in TAC) gave birth to measures like Nouveau-ROUGE [27], that take into account previous summaries to measure redundancy.
- the power of evaluation measures to distinguish consistently between "good" summarizers (usually human) and "bad" or "mediocre" summarizers (usually automatic methods), even across corpora [28]. Rankel et al. [6] use a variety of statistical features from the texts to create a regression-based prediction model that can assign a grade to a given summarization system.
- the lack of completely unsupervised methods (without gold standard summaries) for the evaluation of peer summaries. These methods solving this problem [29, 30] rely on statistical analysis of the content of summaries (term distribution), as well as the source documents to determine the quality of a summary.
- the lack of complementary evaluation measures, that can provide information about different aspects of summary quality (e.g., see [8, 31]).

For an overview of recent summarization evaluation efforts, please also consult [32, Section 5].

In this work, given the numerous efforts on summary evaluation, we study and provide an answer to the following questions:

- Can the combination of existing evaluation measures allow the creation of improved ones, with minor changes? Previous work has shown that some improvement can be achieved by adding liguistic quality information [5] or redundancy checking [27]. Can we achieve a good combination with simply surface-based measures (i.e., minimal preprocessing and no linguistic features)?
- If so, how should we combine these measures?
- Can we use different combinations of methods to grade different aspects of summaries?

In order to answer these questions we combine the well-established n-gram graph methods, under a machine learning perspective. In doing so, we use the individual evaluations as features that describe a single summary and apply regression to model how n-gram graph evaluations can be combined to form the final grade of a summary. We, thus, create a second-level grade estimator (in contrast e.g., to [33]) built as a regression problem, estimating a target grade (e.g., responsiveness or Pyramid score) based on the primary evaluation scores of different methods. We specifically focus on the n-gram graph based approaches (AutoSummENG and MeMoG) due to their purely statistical and language agnostic nature.

3 NPowER: N-Gram Graph Powered Evaluation via Regression

Our method is based on the following simple idea: if there exist a number of rather good grading systems for summaries and these grading systems are not always in agreement, it makes sense to supply an independent judge that can combine the graders' individual estimates to provide a better estimate on the final grade (see also [5]).

In our case we want to determine whether only using surface methods, based on n-gram graphs, we can estimate well-enough (i.e., with a strong correlation to humans) the grades of individual summaries. This is essentially a stronger requirement than that of correlating over whole systems. This is due to the fact that we judge a system based on the average of all its summaries. In fact, we can judge a system well even by taking turns in underestimating it and overestimating it in different summaries, due to the averaging effect. In the case where we judge single summaries this cannot happen.

To answer the questions posed in the previous section we build upon the notion of regression from the domain of statistics and machine learning.

Given a vector of descriptive (independent) features $\overline{x} \in \overline{X}$ and a target (dependent) numeric feature $y \in \mathbb{R}, y = f(\overline{x})$, with f unknown, we want to estimate a (combination) function

$$\tilde{f} : \overline{X} \to \mathbb{R} : \sum (\tilde{f}(\overline{x}) - f(\overline{x}))^2 \to 0, \forall \overline{x} \in X$$

of the descriptive features to best estimate the target feature. In the machine learning literature we find a variety of methods for regression ranging from simple linear regression, to logistic regression (see e.g., [34]) to Support Vector Regression (e.g., [35]). We use a linear regression, where features included in the regression model are selected based on the Akaike Informaion Criterion (AIC) [36], which selects features that best help the estimation without adding too much complexity. The implementation of the linear regression was provided as "Linear Regression" in the WEKA machine learning package [37, Version 3.7].

In the summary evaluation case we consider that the automatic evaluation methods of summaries consist good, descriptive features $\overline{x}$. The target feature y is the manual, human assigned grade. Since in summarization evaluation there exists a variety of human assigned grades, such as responsiveness or Pyramid score, we will need different applications of regression per case.

We examine three different approaches to see whether it makes sense to combine lots of evaluation methods or few, carefully selected ones:

- All: In this case a big set of automatic evaluations (submitted in the AESOP task of Text Analysis Conference) are used as $\overline{x}$ features.
- Only baselines: Only baseline systems are used as $\overline{x}$ features. We consider baselines systems which have been widely used for summary evaluation (i.e., ROUGE-based and BE evaluation).
- Only n-gram graph based: Only the proposed combination of methods is used, namely AutoSummENG and MeMoG, keeping the language-neutral approach of analysis.

We name the application of (linear) regression on the output scores of n-gram graphs methods the *NPowER* method: N-gram graph Powered Evaluation via Regression. We show in following sections that it constitutes a robust, high performing method for summary evaluation, even at the summary level.

Furthermore, we study which $\overline{x}$ features are the most informative, using feature selection, by viewing the evaluation problem as a classification problem. In the following section we report on the experiments and corresponding findings.

4 Experimental Setting and Results

In this section, we describe the data used for the evaluation of the NPowER method and we study how different features contribute to the performance of the system.

4.1 Data

We use the data generated within the AESOP task of the Text Analysis Conferences of 2009 and 2010 (TAC 2009 and TAC 2010[2]). The summaries in the AESOP test data of TAC 2009 consist of all the model summaries and "peer"

[2] See `http://www.nist.gov/tac` (Last visit: Dec 20, 2012) for more information.

(automated, non-model) summaries produced within the TAC 2009 Update Summarization task. 8 human summarizers produced a total of 352 model summaries, and 55 automated summarizers produced a total of 4840 peer summaries. The set of systems included three baseline summarizers. The summaries are split into Initial Summaries (Set A) and Update Summaries (Set B). Update summaries are supposed to take into account the corresponding initial summary and not repeat information on a given topic. In 2009 a total of 12 participants submitted 35 different AESOP metrics, in addition to 2 baselines.

In the 2010 Guided Summarization task, 8 human summarizers produced a total of 368 model summaries, and 43 automatic summarizers produced a total of 3956 automatic summaries. The summaries are split into Main (or Initial) Summaries (Set A) and Update Summaries (Set B), according to the part of the Guided Summarization Task they fall into[3]. Two baseline summarizers were included in the set of automatic summarizers. In 2010 a total of 9 participants submitted 27 different AESOP metrics, in addition to 3 baselines.

The AESOP task is "to create an automatic scoring metric for summaries, that would correlate highly with two manual methods of evaluating summaries, as applied in the TAC 2010 Guided Summarization task" (see TAC 2010 task description), namely the Pyramid method (modified pyramid score) [19] and Overall Responsiveness (see [4]). The scoring metrics (better "measures") are to evaluate summaries including both model (i.e., human generated) and peer (non-model) summaries, produced within the TAC 2010 Guided Summarization task. We note that there were some differences between the 2009 and 2010 datasets:

- In 2009 only ROUGE-SU4 and BE were used as baselines, while ROUGE-2 was added in 2010. In order to provide comparable results across datasets we omitted the ROUGE-2 metric when judging the performance of combined baselines. However, we did not remove it in the cases where all systems were combined ("All" case in the tables of the following section). Thus, we considered it another competing system for the purposes of the experiments.
- The responsiveness grade in 2009 was from 1 to 5, while in 2010 from 1 to 10.

In our experiments we used the TAC provided data for the AESOP task (files: "aesop_allpeers_[A—B]", "manual_allpeers_[A—B]")[4]. We combined the per summary data, by aligning manual grades to their corresponding automatic data lines. In our resulting data, each line contained the following fields:

- Pyramid and Responsiveness scores.
- AESOP SystemID and Topic.
- Evaluation results from each AESOP system.

Below we elaborate on the measures we used, in accordance to current literature, to determine the performance of the evaluation systems we propose.

[3] See `http://www.nist.gov/tac` (Last visit: Dec 20, 2012) for more info on the Guided Summarization Task of TAC 2010.

[4] The data are provided by NIST on request.

4.2 Measuring Correlation – Evaluation Method Performance

In the automatic evaluation of summarization systems we require automatic grades to correlate to human grades. The measurement of correlation between two variables provides an indication of whether two variables are independent or not. Highly correlated variables are dependent on each other, often through a linear relationship. There are various types of correlation measures, called *correlation coefficients*, depending on the context they can be applied. Three types of correlation will be briefly presented here, as they are related to the task at hand:

- The Pearson's product moment correlation coefficient reflects the degree of linear relationship between two variables[5]. The value of Pearson's product moment correlation coefficient ranges from -1 to 1, where 1 indicates perfect positive correlation and -1 perfect negative correlation. Perfect positive correlation indicates that there is a linear relationship between the two variables and that when one of the variables increases, so does the other in a proportional manner. In the case of negative correlation, when one of the two variables increases, the other decreases. A value of zero in Pearson's product moment correlation coefficient indicates that there is no *obvious* correlation between the values of two variables.
- The Spearman's rank correlation coefficient [38] performs a correlation measurement over the ranks of values that have been ranked before the measurement. In other words, it calculates the Pearson's product moment correlation of the ranking of the values of two variables. If two rankings are identical, then the Spearman's rank correlation coefficient will amount to 1. If they are reverse to each other, then the correlation coefficient will be -1. A value of zero in Spearman's rank correlation coefficient indicates that there is no obvious correlation between the rankings of values of two variables. It is important to note that this coefficient type does not assume linear relation between the values, as it uses rankings.
- The Kendall's tau correlation coefficient [39] relaxes one more limitation of the previous methods: it does not expect subsequent ranks to indicate equal distance between the corresponding values of the measured variable.

The above correlation coefficients have all been used as indicators of performance for summary systems evaluation (see, e.g., [1, 15]). To clarify how this happens, consider the case where an *automatic evaluation method* is applied on a set of summarization systems, providing a quantitative estimation of their performance by means of a grade. Let us say that we have assigned a number of *humans* to the task of grading the performance of the same systems as well. If the grades appointed by the method correlate strongly to the grades appointed by humans, then we consider the evaluation method good.

[5] The linear relationship of two correlated variables can be found using methods like linear regression.

4.3 Results — Correlation to Manual Measures

In this first experiment we try the three different sets of $\overline{x}$ features, to determine how well each individual set can perform. We stress that the evaluation we perform is *per summary*. We do this to go more in depth and see whether we can predict the quality of a single summary. If so, we will be able to use the resulting measure as an optimization factor when generating summaries (which is not possible when you have a per system evaluation).

To combine measures we used the WEKA software, as indicated in Section 3. We removed the fields of SystemID and Topic and performed 10-fold cross-validation, using as target variable the corresponding human assigned grade. We used the output file provided by the software as input to the R software [40] and applied correlation tests (cor.test command) between the estimated and true values of the grades.

In Table 1 we show the results of combination, and also provide the performance of the individual baselines (no combination) for reference. We judge performance by all measures of correlation to the human Responsiveness grading. We note that, in all the tables below, the statistical significance p-value of the correlation tests is much lower than 0.001. In the tables below the combination of n-gram graph methods is described as *NPowER*.

In Table 2 we judge performance by all measures of correlation to the human Pyramid grading.

The results of the experiments show the following:

- By combining measures one can significantly improve the estimation of a summary grade, regardless of the underlying measure (responsiveness or Pyramid in our case).
- It appears that using all the measures of AESOP as features, we get the best results in all the cases. However, it might prove impossible to combine all evaluations in a timely manner, since each evaluation represents a completely different system run.

Table 1. Per summary correlation of evaluation measures to Responsiveness

	Setting	Set A			Set B		
Year	$\overline{x}$ features	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
	Baseline: ROUGE-SU4	0.33	0.30	0.22	0.39	0.40	0.29
	Baseline: BE	0.26	0.27	0.20	0.33	0.37	0.27
2009	Baseline comb.	0.34	0.29	0.21	0.39	0.39	0.29
	NPowER	0.60	0.42	0.32	0.61	0.50	0.38
	All	0.83	0.80	0.65	0.67	0.57	0.43
	Baseline: ROUGE-SU4	0.61	0.61	0.47	0.48	0.48	0.38
	Baseline: BE	0.48	0.50	0.38	0.45	0.47	0.36
2010	Baseline comb.	0.61	0.60	0.47	0.50	0.50	0.39
	NPowER	0.72	0.68	0.54	0.73	0.59	0.47
	All	0.75	0.72	0.58	0.74	0.62	0.50

Table 2. Per summary correlation of evaluation measures to Pyramid score

Setting		Set A			Set B		
Year	$\overline{x}$ features	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
2009	Baseline: ROUGE-SU4	0.64	0.65	0.47	0.62	0.60	0.43
	Baseline: BE	0.55	0.58	0.41	0.57	0.59	0.42
	Baseline comb.	0.64	0.65	0.47	0.62	0.62	0.45
	NPowER	0.80	0.73	0.55	0.77	0.69	0.51
	All	0.84	0.79	0.61	0.81	0.76	0.58
2010	Baseline: ROUGE-SU4	0.70	0.72	0.53	0.60	0.62	0.44
	Baseline: BE	0.61	0.64	0.46	0.55	0.58	0.42
	Baseline comb.	0.71	0.73	0.54	0.61	0.64	0.47
	NPowER	0.83	0.80	0.61	0.79	0.72	0.54
	All	0.85	0.83	0.64	0.81	0.75	0.56

- By only using baseline combination we do better than by using individual baselines; but not much better in most cases.
- By only using n-gram graph based methods combined (AutoSummENG and MeMoG in our case) we can significantly outperform the combination of baselines and even approach the performance of using all the systems (in most cases). In other words, 2 measures combined are performing close to more than 20 measures combined.

Using only the n-gram graph systems we performed another experiment to see how transferable the learnt regression models are across summary groups or different data (years):

- In the first experiment we train the regression model with all the 2009 data (from both sets) and test the model on the 2010 data (both sets). Then, we switch training and test sets and repeat the experiment. We describe this experiment as the "across years" experiment.
- In the second experiment we train the regression model with all the Set A data (from both 2009 and 2010) and test the model on all the Set B data (from both 2009 and 2010). Then, we switch training and test sets and repeat the experiment. We describe this experiment as the "across sets" experiment.

We illustrate the results of both experiments on Table 3. The experiment across years for Responsiveness offers good results, since the correlation scores remain high in both cases. We remind the reader that the correlation is judged on a per summary basis, which means that the per system performance is expected to be higher. The experiment across sets is equally interesting and promising. We see that the results are still high and the performance is almost identical (when rounding to the second deciman) in the two different settings. Of course, to be able to judge the robustness of the method with certainty, more experiments must be run (starting possibly from sampling the existing datasets). However, these first results are indications of acceptable stability on the examined tasks.

Table 3. Correlations between NPowER grades and Responsiveness (left), Pyramid (right) across years (top half) and sets (bottom half)

Setting		Target: Responsiveness			Target: Pyramid score		
Train Year	Test Year	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
2009	2010	0.72	0.65	0.52	0.72	0.74	0.55
2010	2009	0.61	0.47	0.35	0.78	0.72	0.74
Setting		Target: Responsiveness			Target: Pyramid score		
Train Set	Test Set	Pearson	Spearman	Kendall	Pearson	Spearman	Kendall
A	B	0.64	0.55	0.42	0.75	0.69	0.51
B	A	0.63	0.55	0.42	0.76	0.71	0.53

The results while optimizing for Pyramid scores were even better, as we illustrate on the right-hand side of Table 3. Overall, the method appears to be very effective across sets and years, forming a very interesting and useful estimator of summary quality.

On the other hand, deeper analysis shows that NPowER is not a perfect measure. This is clearly shown from the Kendall's tau value: an ideal measure that would indicate which summary of a pair is better, like a human, would have a value very close to 1.0. Thus, there is still much space for improvement. But, how can we improve? Can we determine which features are important and which are missing? To start addressing these questions we perform a feature study in the following section.

4.4 Results — Feature Selection

In this section, we examine which features are most informative for the estimation of a responsiveness grade of a system. In order to be able to apply information theory methods, such as Information Gain on the $\overline{x}$ features, we consider the grading problem as a *classification problem*. In the case of responsiveness we have 10 different possible classes, one per assignable grade (from 1 to 10).

The Information Gain (IG) measure is a measure of how "predictive" of a class a single feature is: $IG(\text{Class}, \text{Attribute}) = H(\text{Class}) - H(\text{Class}|\text{Attribute})$, where $H(x)$ is the entropy of the x values, and $H(x|y)$ is the entropy of x given y.

We use the 2009 and 2010 datasets Set A. We only focus on set A, because for set B (update task) information from set A should be used and we are trying to avoid inter-set dependecies at this point. The features with the top 10 information gain (IG) values are as follows:

The drawback of the IG measure is that it judges one feature at a time and does not offer combination information. Furthermore, by converting the regression problem to a classification problem we apply the same penalty to grades estimations that are not on-target, regardless of how different the grade was from the target value. However, it provides a hint at which features are more likely to help when determining the right grade: the n-gram graph features are

Table 4. Top 10 Information Gain features on TAC 2009 Responsiveness score classification problem

Set A			
2009		2010	
IG	System	IG	System
0.3852	MeMoG	0.5446	MeMoG
0.3814	AutoSummENG	0.5388	S7
0.3131	S17	0.4644	S9
0.3129	S22	0.453	S21
0.2989	S12	0.4497	AutoSummENG
0.2984	S20	0.4357	S17
0.2562	S19	0.4155	S18
0.2561	S14	0.4094	S10
0.2485	S16	0.4052	ROUGESU4
0.2425	S18	0.3472	S12

consistently highly graded. It is also noticeable that ROUGESU4 is within the table for the case of 2010, illustrating that baselines are important.

In order to see why combining methods offers additional information we studied the correlation between the baselines and the n-gram graph methods within NPowER. The results showed that the features were not too strongly correlated (0.70 Pearson correlation). We believe that the fact that each of the methods is correlated to the target features (e.g., responsiveness), but they are not highly correlated to each other makes their combination useful. It would make sense to determine, ideally orthogonal, evaluation measures to maximize the combination effect. Equivalently, it would make sense to *analyze human answers to orthogonal axes*, which would in turn be estimated by automatic measures.

5 Conclusion

In this paper we proposed a novel summary evaluation method, based on the linear combination of surface (n-gram graph) methods. The method is termed NPowER. We showed that combining several measures improves the estimation of summary quality. We then showed that NPowER is highly competitive when aiming to estimate two different, human evaluation measures (responsiveness and Pyramid score) on the summary level. We briefly studied the improtance of evaluation measures in term of information theory, by viewing the grading of a summary as a classification problem.

Our study showed that combining measures can prove effective, but there is significant space for improvement if we want to be able to confidently judge a summary automatically. Our future aims are to see whether combining state-of-the-art methods covering a variety of qualitative aspects (such as linguistic quality and coherence). We aim to examine whether these aspects are indeed uncorrelated enough to provide complementary information towards the best

evaluation possible. We furthermore will try to examine, by viewing the evaluation process as a classification process, how one can improve the performance by viewing each grade as a different class.

References

[1] Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), pp. 25–26 (2004)
[2] Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics (2002)
[3] Dang, H.T.: Overview of DUC 2005. In: Proceedings of the Document Understanding Conf. Wksp. (DUC 2005) at the Human Language Technology Conf./Conf. on Empirical Methods in Natural Language Processing, HLT/EMNLP 2005 (2005)
[4] Dang, H.T., Owczarzak, K.: Overview of the TAC 2008 update summarization task. In: TAC 2008 Workshop - Notebook Papers and Results, Maryland MD, USA, pp. 10–23 (2008)
[5] Conroy, J.M., Dang, H.T.: Mind the gap: Dangers of divorcing evaluations of summary content from linguistic quality. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester, UK, Coling 2008 Organizing Committee, pp. 145–152 (2008)
[6] Rankel, P., Conroy, J., Schlesinger, J.: Better metrics to automatically predict the quality of a text summary. Algorithms 5, 398–420 (2012)
[7] Giannakopoulos, G., El-Haj, M., Favre, B., Litvak, M., Steinberger, J., Varma, V.: TAC 2011 MultiLing pilot overview. In: TAC 2011 Workshop, Maryland MD, USA (2011)
[8] Owczarzak, K., Conroy, J., Dang, H., Nenkova, A.: An assessment of the accuracy of automatic evaluation in summarization. In: NAACL-HLT 2012, p. 1 (2012)
[9] Mani, I., Bloedorn, E.: Multi-document summarization by graph search and matching. In: Proceedings of AAAI 1997, pp. 622–628. AAAI (1997)
[10] Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study: Final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, vol. 1998 (1998)
[11] Van Halteren, H., Teufel, S.: Examining the consensus between human summaries: Initial experiments with factoid analysis. In: Proceedings of the HLT-NAACL 2003 on Text Summarization Workshop, vol. 5, pp. 57–64. Association for Computational Linguistics, Morristown (2003)
[12] Lin, C.Y., Hovy, E.: Manual and automatic evaluation of summaries. In: Proceedings of the ACL 2002 Workshop on Automatic Summarization, vol. 4, pp. 45–51. Association for Computational Linguistics, Morristown (2002)
[13] Jones, K.S.: Automatic summarising: The state of the art. Information Processing & Management, Text Summarization 43, 1449–1481 (2007)
[14] Baldwin, B., Donaway, R., Hovy, E., Liddy, E., Mani, I., Marcu, D., McKeown, K., Mittal, V., Moens, M., Radev, D.: Others: An evaluation roadmap for summarization research. Technical report (2000)
[15] Nenkova, A.: Understanding the Process of Multi-Document Summarization: Content Selection, Rewriting and Evaluation. PhD thesis (2006)

[16] Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In: ANLP/NAACL Workshop on Summarization (2000)
[17] Marcu, D.: Theory and Practice of Discourse Parsing and Summarization, The. The MIT Press (2000)
[18] Saggion, H., Lapalme, G.: Generating indicative-informative summaries with sumum. Computational Linguistics 28, 497–526 (2002)
[19] Passonneau, R.J., McKeown, K., Sigelman, S., Goodkind, A.: Applying the pyramid method in the 2006 document understanding conference. In: Proceedings of Document Understanding Conference (DUC) Workshop 2006 (2006)
[20] Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311–318 (2001)
[21] Hovy, E., Lin, C.Y., Zhou, L., Fukumoto, J.: Basic elements (2005)
[22] Hovy, E., Lin, C.Y., Zhou, L., Fukumoto, J.: Automated summarization evaluation with basic elements. In: Proceedings of the Fifth Conference on Language Resources and Evaluation, LREC (2006)
[23] Owczarzak, K.: Depeval (summ): dependency-based evaluation for automatic summaries. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 1, pp. 190–198. Association for Computational Linguistics (2009)
[24] Giannakopoulos, G., Karkaletsis, V., Vouros, G., Stamatopoulos, P.: Summarization system evaluation revisited: N-gram graphs. ACM Trans. Speech Lang. Process. 5, 1–39 (2008)
[25] Giannakopoulos, G., Karkaletsis, V.: Summarization system evaluation variations based on n-gram graphs. In: TAC 2010 Workshop, Maryland MD, USA (2010)
[26] Schilder, F., Kondadadi, R.: A metric for automatically evaluating coherent summaries via context chains. In: IEEE International Conference on Semantic Computing, ICSC 2009, pp. 65–70 (2009)
[27] Conroy, J., Schlesinger, J., O'Leary, D.: Nouveau-rouge: A novelty metric for update summarization. Computational Linguistics 37, 1–8 (2011)
[28] Amigó, E., Gonzalo, J., Verdejo, F.: The heterogeneity principle in evaluation measures for automatic summarization. In: Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, pp. 36–43. Association for Computational Linguistics, Stroudsburg (2012)
[29] Louis, A., Nenkova, A.: Automatically evaluating content selection in summarization without human models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 1, pp. 306–314. Association for Computational Linguistics (2009)
[30] Saggion, H., Torres-Moreno, J., Cunha, I., SanJuan, E.: Multilingual summarization evaluation without human models. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 1059–1067. Association for Computational Linguistics (2010)
[31] Vadlapudi, R., Katragadda, R.: Quantitative evaluation of grammaticality of summaries. In: Gelbukh, A. (ed.) CICLing 2010. LNCS, vol. 6008, pp. 736–747. Springer, Heidelberg (2010)
[32] Lloret, E., Palomar, M.: Text summarisation in progress: a literature review. Artificial Intelligence Review (2011)

[33] Pitler, E., Louis, A., Nenkova, A.: Automatic evaluation of linguistic quality in multi-document summarization. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 544–554. Association for Computational Linguistics (2010)
[34] Menard, S.: Applied logistic regression analysis, vol. 106. Sage Publications, Incorporated (2001)
[35] Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. Software 80, 604–611 (2001), http://www.Csie.Ntu.Edu.Tw/cjlin/libsvm
[36] Akaike, H.: Likelihood of a model and information criteria. Journal of Econometrics 16, 3–14 (1981)
[37] Witten, I.H., Frank, E., Trigg, L., Hall, M., Holmes, G., Cunningham, S.J.: Weka: Practical machine learning tools and techniques with java implementations. In: ICONIP/ANZIIS/ANNES, pp. 192–196 (1999)
[38] Spearman, C.: Footrule for measuring correlation. British Journal of Psychology 2, 89–108 (1906)
[39] Kendall, M.G.: Rank Correlation Methods. Hafner New York (1962)
[40] Team, R.C.: R: A Language and Environment for Statistical Computing. In: R Foundation for Statistical Computing, Vienna, Austria (2012) ISBN 3-900051-07-0