A Repository of State of the Art and Competitive Baseline Summaries for Generic News Summarization

Kai Hong[1], John M. Conroy[2], Benoit Favre[3], Alex Kulesza[4], Hui Lin[5], Ani Nenkova[1]

[1]University of Pennsylvania, Philadelphia, PA, USA
[2]IDA/Center for Computing Sciences, Bowie, Maryland, USA
[3]Aix-Marseille Univ LIF/CNRS, Marseille, France
[4]University of Michigan, Ann Arbor, MI, USA
[5]LingoChamp, Shanghai, China

[1]{hongkai1,nenkova}@seas.upenn.edu, [2]conroy@super.org
[3]benoit.favre@lif.univ-mrs.fr, [4]kulesza@umich.edu, [5]h@liulishuo.com

Abstract

In the period since 2004, many novel sophisticated approaches for generic multi-document summarization have been developed. Intuitive simple approaches have also been shown to perform unexpectedly well for the task. Yet it is practically impossible to compare the existing approaches directly, because systems have been evaluated on different datasets, with different evaluation measures, against different sets of comparison systems. Here we present a corpus of summaries produced by several state-of-the-art extractive summarization systems or by popular baseline systems. The inputs come from the 2004 DUC evaluation, the latest year in which generic summarization was addressed in a shared task. We use the same settings for ROUGE automatic evaluation to compare the systems directly and analyze the statistical significance of the differences in performance. We show that in terms of average scores the state-of-the-art systems appear similar but that in fact they produce very different summaries. Our corpus will facilitate future research on generic summarization and motivates the need for development of more sensitive evaluation measures and for approaches to system combination in summarization.

Keywords: generic summarization, DUC 2004, evaluation

1. Introduction

In generic multi-document news summarization, a system is expected to produce a summary of the most important information conveyed in a set of topically related news articles. The task does not presuppose the existence of a user query or information need (thus the name generic), so it appears to call for an automatic method of at least minimal semantic processing to discover important information, a problem with huge appeal for cognitive science and artificial intelligence enthusiasts. Not surprisingly, many competing approaches have been developed for the task. The early editions of the Document Understanding Conference (DUC) evaluations (Over et al., 2007), carried out between 2001 and 2004, have had a strong formative influence on work on generic multi-document news summarization, by providing annual test collections for a shared task. In later years of evaluation the task was abandoned, in favor of other summarization problems such as topic-focused summarization in response to a user need, summarizing given the readers' prior knowledge of the topic or summarizing in specific domains where the aspects of importance are clearly defined a priori.

Much of the scientific literature on summarization, however, remained focused on generic summarization, rendering the DUC 2004 (Task 2) dataset the most appropriate one for comparing the systems. The dataset consists of 50 inputs for multi-document summarization, each consisting of about 10 news articles related to a topic. The task was to generate a 100-word summary for each input.

Some recent systems have been evaluated on DUC 2004, some have been evaluated on earlier datasets and some on data for topic-focused summarization, where the summarizers simply ignore the topic. Even those evaluated on the DUC'04 dataset cannot be directly compared in most cases. The reason for this confusion is the number of parameters one can set for ROUGE (Lin, 2004), the automatic summarization evaluation of choice. ROUGE evaluates summaries by comparing the n-gram overlap between a summary and a set of gold-standard summaries produced by people. Obviously the choice of n-gram size—1-, 2-, 3-, 4-, skip bigram or basic syntactic elements—would change the score. Furthermore the tool produces recall, precision and f-measure results, so different researchers have reported a different combination of the above. To make matters even more disorienting, the overlap with the gold-standard can be computed with function words preserved or excluded and with words stemmed or not. The system summary can be left as is, truncated to 100 words as required in the original task definition of DUC 2003 as well as TAC 2008-2011, or to 665 bytes as in the original task definition of DUC 2004, resulting in variable length summaries.

System Peer 65 (CLASSY 04), the best official entry in the DUC 2004 evaluation, is a specific case in point. A quick survey of the literature shows that its ROUGE-1 recall has been reported as 39.1% (Manna et al., 2012), 38.3% (Lin and Bilmes, 2011), and 30.8% (Conroy et al., 2006), depending on the parameters used for evaluation.

The repository we introduce in this paper will remove some of the stumbling blocks in evaluation and comparison of generic summarization systems. We provide the output of several state-of-the-art summarizers on DUC 2004. For completeness of the comparison we also include several popular baselines[1].

We then adopt the recommendations from recent findings on summarization evaluation (Owczarzak et al., 2012) to fix the ROUGE parameters to those that lead to highest agreement with manual evaluation. We report three ROUGE scores which have been shown to have good and complementary behavior. We perform paired tests for statistical significance to establish the superiority of one system over another (Rankel et al., 2011) by abstracting away the noise of intrinsic input difficulty (Nenkova and Louis, 2008).

We find that according to ROUGE-2 recall, the best systems developed after CLASSY 04 are in fact not significantly different from each other (Nenkova and Louis, 2008).

We further analyze one of the possible reasons for the similar performance of the top systems. It is conceivable that despite using markedly different approaches to select important content, the systems may end up choosing the same facts or even the same sentences. To quantify the extent to which this overlap in selection choices occurs, we compare the summaries produced by the three best systems at three levels of granularity: sentences, words and summary content units (SCU) (Nenkova et al., 2007). At all three granularities the overlap in content across summaries is low. Our findings suggest that summarization research may benefit from the development of more sensitive measures of content to capture finer nuances in system performance. They also reveal there is a high potential for system combination in summarization.

Below we first describe the systems whose output we include in our repository. We start with a description of the baselines, then present the state-of-the-art systems. In the next section we present ROUGE-1, -2 and -4 recall scores produced with the same setting of ROUGE to finally enable direct comparison of the systems. Then we present results on statistical significance in the differences between systems. Finally we investigate the overlap between summaries from our collection of the state-of-the-art systems.

2. Baseline Systems

Word probability (FreqSum): A simple yet powerful approach for multi-document news summarization is to approximate the importance of words with their probability in the input, then select sentences with high average word probability (Nenkova et al., 2006). We follow this approach, called FreqSum, changing only the way we handle redundancy in the summary. We do not include a sentence in the summary if its cosine similarity with a sentence already in the summary exceeds a predefined threshold determined on development data (DUC 2003).

We use the same approach for handling redundancy for the other summarizers—TsSum, Centroid, Cont. LexRank, and RegSum—which rank sentences according to other criteria. We describe these next.

Topic words (TsSum): Another powerful method of weighting words is the application of the log-likelihood ratio (LLR) test, usually called topic signatures, which compares the distribution of words in the input and a large background corpus (Lin and Hovy, 2000). Since the log-likelihood ratio λ follows a χ-square distribution, it provides a natural way to select topic words by setting a predefined confidence level cutoff. Words with scores above the threshold are topic words with equal importance (their weight is set to 1). Sentence scores are computed as the ratio of unique topic words to the total number of unique words in the sentence, as suggested by Conroy et al., (2006). Also following their work we consider words to be topic words if their χ-square statistic exceeds 10, corresponding to a 99.9% confidence level.

Centroid: The centroid of a document is an abstract sentence, containing all the words in the document. The weight of each word in the centroid is the average tf-idf weight of the word in all input sentences, assuming words have zero weight in sentences in which they do not appear. The centroid score for the sentences in the input is calculated as their cosine similarity with the centroid (Radev et al., 2004b). We compute these scores from the popular MEAD system directly (Radev et al., 2004a). Note the MEAD system scores sentence importance as a linear combination of the centroid score and the position score of the sentence in the document. We do not use position information here.

Cont. LexRank: Graph approaches (Mihalcea and Tarau, 2004; Wan and Yang, 2008) are widely used for summarization. LexRank (Erkan and Radev, 2004) is arguably the most popular member of this class of summarizers. The input text is represented as a graph $G(V, E)$, where V is the set of sentences in the input. There is an edge e_{ij} between two nodes v_i and v_j if and only if the cosine similarity between them is above a certain threshold. Since there could be multiple possibilities of choosing the threshold in Lexrank, we here employ continuous Lexrank instead (Erkan and Radev, 2004). The cosine similarity between v_i and v_j is directly used as weight for e_{ij}. Sentence importance is calculated by running the PageRank (Lawrence et al., 1998) algorithm on the text graph. We compute the sentence importance from MEAD directly.

Greedy-KL: This method is related to the word probability approach. Here, however, the goal is to incorporate knowledge about the overall distribution of words in the input, instead of focusing on the probability of individual words. The idea is to minimize the KL divergence between the probability distribution of words estimated from the summary and that from the input. Since finding the summary with smallest KL divergence is intractable, we present results for the greedy KL summarizer, as described by Haghighi and Vanderwende (2009). The approach iteratively selects the next sentence s_i to be included in the summary C, which is done by picking the sentence that will minimize the KL divergence between the word distributions

[1]The corpus is publicly available: https://www.seas.upenn.edu/~nlp/corpora/sumrepo.html

estimated from $S = s_i \bigcup C$ and the original input D [2].

3. State-of-the-art Systems

We present the state-of-the-art systems in this section. Peer 65 is introduced first, followed by the other systems presented in alphabetical order.

CLASSY 04 [Peer 65]: This system (Conroy et al., 2004) was the best among those that entered the official DUC 2004 evaluation. It is often used as comparison system by developers of novel summarization methods. It employs a Hidden Markov Model, using topic signature as the only feature. The probability of one sentence being selected in the summary also depends on the importance assigned to its adjacent sentences in the input document. It is worth noting that there is a linguistic preprocessing component in this system.

CLASSY 11: This is the successor of Peer 65, developed to handle query-focused summarization. CLASSY 11 was the best query-focused system according to one of the manual evaluation measures (overall responsiveness) in the official TAC 2011 evaluation (Conroy et al., 2011). Like CLASSY 04, it uses topic signatures as features; however, it attempts to estimate the probability that a term (bigram) will occur in a human-generated summary. A subset of non-redundant sentences with highest scores is selected using non-negative matrix factorization algorithm. Two major changes to CLASSY 11 were made for this study. First, for ease of comparison with CLASSY 04 (Peer 65), the 2004 linguistic preprocessing was used. Likewise, a generic summarization term-weighting method was needed, so the LSA based approach of OCCAMS_V (which we will describe afterwards) was deployed.

DPP: Determinantal point processes (DPPs) (Kulesza and Taskar, 2012) are probabilistic models of sets which balance the selection of important information and diverse groups of sentences within a given length. Specifically, DPPs combine a per-sentence quality model that prefers relevant sentences with a global diversity model encouraging non-overlapping content. This setup has several advantages. First, by treating these opposing objectives probabilistically, there is a rigorous framework for trading off between them. Second, the sentence quality model can depend on arbitrary features, and its parameters can be efficiently learned from reference summaries via maximum likelihood training; in contrast, most standard summarization techniques are tuned by hand. Finally, because a DPP is a probabilistic model, at test time it is possible to sample multiple summaries and apply minimum Bayes risk decoding, thus improving ROUGE scores. The DPP model in this work is trained on the DUC 2003 data to optimize the ROUGE-1 F-score.

ICSISumm: The ICSI summarization system adopts a global linear optimization framework, finding the globally optimal summary rather than greedily choosing sentences according to their importance. A summary is generated by covering the most important concepts in the document set. Even though Integer Linear Programming (ILP) is NP-hard, the exact solutions to this problem can still usually be found by a standard ILP solver in a very fast fashion (Gillick and Favre, 2009). We collect the summaries for DUC 2004 generated by the ICSI/UTD summarization system at TAC 2009 (Gillick et al., 2009), which optimizes the coverage of key bigrams weighted by their frequency in the document collection.

OCCAMS_V: This system (Davis et al., 2012; Conroy et al., 2013) employs latent semantic analysis (LSA) to compute term weights and a sentence selection algorithm based on two combinatorial problems, the budgeted maximal coverage (BMC) problem and the knapsack problem. The sentence selection algorithm extends the work of Samir Khuller on BMC (Khuller et al., 1999), who proposed a $(1 - e^{-\frac{1}{2}})$ approximation algorithm, whereas a greedy algorithm has an unbounded approximation ratio. OCCAMS_V improves the algorithm by applying a fully polynomial-time approximation scheme (FPTAS) dynamic programming algorithm to knapsack problems formed from subsets of candidate sentences found by applying greedy BMC with a larger bound (e.g., word counts). As was done with CLASSY 11, the CLASSY 04 linguistic preprocessing was used for OCCAMS_V.

RegSum: The RegSum system (Hong and Nenkova, 2014) employs a supervised model for predicting word importance. This model is superior to prior methods for identifying the words which are included in human models. RegSum combines the weights estimated from three unsupervised approaches, along with features including locations, part-of-speech, name-entity-tags, topic categories and contexts. Specifically, this system captures words which are of intrinsic interest to people by analyzing a large number of summary-abstract pairs from the New York Times corpus (Sandhaus, 2008). The summarizer employs the same greedy optimization framework as FreqSum and TsSum. It shows that the quality of the summaries could be greatly improved by better estimation of word importance.

Submodular: Treating multi-document summarization as a submodular maximization problem has proven successful (Lin and Bilmes, 2011) and has spurred a great deal of interest in this line of research (Sipos et al., 2012; Morita et al., 2013; Dasgupta et al., 2013). The advantage of using a submodular function to estimate summary importance is that there is an efficient algorithm for incrementally computing the importance of a summary with a performance guarantee on how close the approximate solution will be to the globally optimal one.

We collect summaries from (Lin and Bilmes, 2012), where they employ structure learning to produce the submodular functions. They first learn a mixture of submodular "shells" in a max-margin structured prediction setting. Then a mixtures of the shells can be instantiated to generate a more complex submodular function.

[2] This is different from the formula in Haghighi and Vanderwende (2009), as we minimize $KL(S \parallel D)$ instead of $KL(D \parallel S)$. The summarizer we present here is the one with better performance.

4. ROUGE Performance

We conduct our ROUGE experiment following the standard suggested by Owczarzak et al. 2012[3], where ROUGE-2 recall with stemming and stopwords not removed provides the best agreement with manual evaluations. We also compute ROUGE-1 recall, which is the measure with highest recall of ability to identify the better summary in a pair, and ROUGE-4 recall, which is the measure with highest precision of ability to identify the better summary in a pair (Owczarzak et al., 2012).

Each summary is truncated to 100 words automatically by ROUGE while evaluation.[4]

System	R-1	R-2	R-4
Cont. LexRank	35.95	7.47	0.82
Centroid	36.41	7.97	1.21
FreqSum	35.30	8.11	1.00
TsSum	35.88	8.15	1.03
Greedy-KL	37.98	8.53	1.26
CLASSY 04	37.62	8.96	1.51
CLASSY 11	37.22	9.20	1.48
Submodular	39.18	9.35	1.39
DPP	**39.79**	9.62	1.57
RegSum	38.57	9.75	1.60
OCCAMS_V	38.50	9.76	1.33
ICSISumm	38.41	**9.78**	**1.73**

Table 1: System performance comparison (%)

Table 1 shows the performance of all approaches according to ROUGE-1,2,4 recall, sorted by ROUGE-2 recall in ascending order. Among the baseline systems, Greedy-KL performs the best according to three ROUGE scores. It even achieves ROUGE-1 recall higher than CLASSY 04 and CLASSY 11. The system with highest ROUGE-1 recall is DPP, which exceeds the R-1 recall of CLASSY 04 by 2.37%. It also achieves better performance on R-1 recall than all of the other systems by at least 1%, except for the Submodular system for which the difference is 0.59%. Note that this system's edge over the other approaches is not so clear-cut on R-2 and R-4, which likely is a reflection of the fact that DPP optimizes R-1 F-score during the learning process. The Integer Linear Programming system (ICSISumm) optimizes the best coverage over bigram frequency, and achieves the highest scores on R-2 and R-4. The systems perform extremely close on ROUGE-2 recall, with the top six systems only differing within a range of 0.58%. CLASSY 04 has a strong performance on ROUGE-4 recall (1.51%), and only three systems (DPP, RegSum, ICSISumm) exhibit better performance on this evaluation.

[3]ROUGE-1.5.5 with the parameters: -n 4 -m -a -l 100 -x -c 95 -r 1000 -f A -p 0.5 -t 0

[4]In the official DUC 2004 evaluation, summaries of length 665 bytes were required. This meant that systems could produce different numbers of words. The variation in length has a noticeable impact on ROUGE scores, especially recall, which is the ROUGE score most highly correlated with manual evaluations.

5. Significance Test

We conduct two-sided Wilcoxon signed-rank tests between each pair of the state-of-the-art systems as well as Greedy-KL, as advocated in Rankel et al. (2011). The experiment is performed on ROUGE-1,2,4. Detailed p-values for the comparison are presented in Table 7, Table 8 and Table 9. The order of systems in these tables are listed according to the order they are described in Section 2. and Section 3. P-values indicating significance at the 95% confidence level or above are shown in bold. A plus sign before the p-value indicates that the system in the row performs better than the one in the column. A minus indicates the opposite relation, that the system in the column is better than the one listed in the row.

On ROUGE-1, only DPP and Submodular show significant improvement over CLASSY 04 and Greedy-KL. DPP performs significantly better than all but one systems (submodular system) on R-1 recall.

On ROUGE-2, there are no significant differences between the top six systems. RegSum is the only one which provides a significant improvement over CLASSY 04 ($p = 0.0483$). The difference between CLASSY 04 and OCCAMS_V and ICSISumm however tends towards significance, with p-values of 0.0572 and 0.071 respectively. Five of the state-of-the-art systems are significantly better than Greedy-KL on R-2.

On ROUGE-4, there are also no recently developed systems that significantly outperform CLASSY 04. ICSISumm has the best overall performance but the p-value for the difference is 0.1798. DPP, ICSISumm and Regsum are significantly better than Greedy-KL on R-4 recall.

6. Overlap Between Summaries

We have so far established that the state-of-the-art summarization systems developed in recent years comfortably outperform standard baselines and also work better than the state-of-the-art of a decade ago. The newer systems however do not appear to be that different from each other in automatic evaluation. In the rest of the paper we examine if the lack of difference between systems is due to the fact that they select the same content, albeit via different methods. We study the overlap of the produced summaries in terms of sentences, words and, for a subset of inputs, in terms of manually annotated summary content units following the Pyramid Method (Nenkova et al., 2007).

6.1. Sentence Level Comparison

Given that all the systems we study are extractive, selecting sentences directly from the input, it is fairly easy to compute sentence overlap of the summaries for the same input produced by different systems. We use the Jaccard coefficient to compute the degree of sentence overlap. If we denote the sets of sentences from two summaries of the same input S as A_S and B_S, the Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the two sets:

$$J(A_S, B_S) = \frac{|A_S \bigcap B_S|}{|A_S \bigcup B_S|}$$

	Greedy-KL	CLASSY 04	CLASSY 11	DPP	ICSISumm	OCCAMS_V	RegSum	Submodular
Greedy-KL		0.084	0.141	0.074	0.098	0.035	0.111	0.135
CLASSY 04			0.003	0.075	0.082	0.003	0.136	0.050
CLASSY 11				0.073	0.092	0.169	0.060	0.070
DPP					0.143	0.071	0.124	0.082
ICSISumm						0.083	0.149	0.038
OCCAMS_V							0.090	0.050
RegSum								0.113
Submodular								

Table 2: Sentence overlap using the Jaccard coefficient

	Greedy-KL	CLASSY 04	CLASSY 11	DPP	ICSISumm	OCCAMS_V	RegSum	Submodular
Greedy-KL		0.287	0.342	0.315	0.353	0.265	0.359	0.348
CLASSY 04			0.206	0.300	0.315	0.208	0.407	0.289
CLASSY 11				0.311	0.295	0.385	0.262	0.317
DPP					0.387	0.307	0.372	0.367
ICSISumm						0.302	0.375	0.307
OCCAMS_V							0.274	0.292
RegSum								0.366
Submodular								

Table 3: Word overlap using the Jaccard coefficient, including stopwords, with stemming

The Jaccard coefficient between two systems is defined to be the mean of $J(A_S, B_S)$ across all inputs S in the DUC 2004 dataset.

Occasionally, some sentences in the summary do not match exactly any of the sentences in the input. This happens for almost all last sentences in the summary which were truncated mid-sentence in order to meet the 100 word limit. We simply ignore the the last sentence if it does not match. In other cases the sentence was not found because the system performed some sentence editing or simplification. For example CLASSY 04, CLASSY 11 and OCCAMS_V delete appositive clauses. In this case we find the most similar sentence in the input measured by cosine similarity and treat that input sentence as the one that appeared in the summary. In Table 2, we show the Jaccard coefficients between each pair of systems.

The degree of overlap at sentence level is surprisingly low. The Jaccard coefficients between two of the best systems, DPP and ICSISumm is 0.143. OCCAMS_V and CLASSY 11 extract the most similar sets of sentences, with a Jaccard coefficient of 0.169. The lowest overlap at the sentence-level is between CLASSY 04 and CLASSY 11, with Jaccard coefficient of only 0.003.

6.2. Word Level Comparison

We next investigate the word overlap between summaries. Specifically, we compute the Jaccard coefficient between the sets of unique words in each summary for a given input. As in the ROUGE setting which correlates the best with human evaluations, we perform stemming and include stopwords while doing the comparison. Table 3 lists the Jaccard coefficients for word overlap between systems. The coefficients range from 0.206 (CLASSY 04 vs CLASSY 11) to 0.407 (CLASSY 04 vs RegSum), with a mean of 0.318. The numbers are low overall, with systems sharing at most a third of the unique words in their summaries.

6.3. SCU Level Comparison

The comparison of sentence and word overlap does not directly reveal the overlap in semantic content expressed in the summaries. To better investigate the coverage of information in summaries, we compute the overlaps of SCUs using the Pyramid Method (Nenkova et al., 2007). SCUs are defined as semantically motivated subsentential units. They are annotated manually and map together all expressions of the same content, even if the wording across summaries differs. First, we manually create the pyramids for 10 input sets in DUC 2004; that is, we find the SCUs expressed in the four human written abstracts for the input which were provided as part of the DUC data. Then we identify which SCUs are expressed in the summaries of four of the automatic systems—Greedy-KL, DPP, ICSISumm and RegSum—for those inputs. SCUs conveyed in truncated last sentences are also included in the calculations of overlap.

On average there are 33.7 SCUs expressed in the four human summaries. On average each SCU is expressed 1.78 times across the four human summaries, with few SCUs repeated frequently and most appearing in only one summary. Table 4 shows the average number of SCUs from the reference human summaries that are also expressed in the machine summaries. The total number of unique SCUs covered by the four machine summaries is 13.6 on average for the 10 inputs. We also show the modified pyramid scores for those four summaries in Table 4. There are no significant differences when evaluated by pyramid scores.

We compute the similarity of systems in terms of SCUs using the Jaccard coefficient. The result is shown in Table 5. The similarity score ranges from 0.347 to 0.426, with an average of 0.385.

	KL	DPP	ICSISumm	RegSum
# SCUs	7.3	6.9	7.1	7.3
Pyramid Score	0.517	0.506	0.514	0.533

Table 4: Average number of SCUs per summary and modified pyramid scores on the first 10 input sets

Even in the manual analysis of content overlap, the systems appear to share some content but more than half of the information appear only in machine summary produced by one of the two systems. Clearly the systems perform similarly not because they end up choosing the same content. The systems choose different but equally good content. This finding indicates that it is quite possible to successfully exploit methods for system combination in order to combine content from each of the systems. This idea is also supported by work on fully automatic evaluation for summarization which has shown that the combination of different systems' input serves as an excellent reference for estimating summary content quality (Louis and Nenkova, 2013).

Alternatively, it may be possible to develop more sensitive evaluation methods that are capable of identifying differences in the importance of non-shared content in the summaries. This development may be necessary for further progress in content selection for summarization, as even the manual evaluations do not find many significant differences between the top systems (Rankel et al., 2013).

	KL	DPP	ICSISumm	RegSum
KL		0.382	0.386	0.426
DPP			0.347	0.419
ICSISumm				0.351
RegSum				

Table 5: SCUs overlap using the Jaccard coefficient

To more concretely illustrate the differences in summary content, we show the machine summaries for input d30002t and d30010t in Table 6. These are the two input with the least and most of overlap between SCUs, with an average Jaccard coefficient of 0.140 and 0.529, respectively. We marked the expressions of SCUs which appear in multiple machine summaries in Table 6. For the input d30002t, 12 unique SCUs are expressed in the machine summaries. Only three of them are expressed in more than one machine summary. SCUs A, B and C are respectively **Hurricane Mitch brought huge death toll in Central America**, **Slow-moving Mitch battered the Honduran for more than a day** and **Taiwan sent aid to Central American countries**. Apart from those SCUs, there are large differences in the content that appear in the summaries. On the other hand, of the 14 unique SCUs from the human summaries on input d30010t, eight of them are expressed in more than one machine summary. It is worth noticing that six of the 14 SCUs appear in at least three machine summaries.

7. Conclusion

We presented a repository of generic multi-document summaries produced by a range of systems for the same input. This resource allow us to carry out a unique comparison between existing approaches. We have also outlined the methodology for reporting results, establishing informed choices for ROUGE settings and for the computation of statistical significance.

We demonstrate that the greedy KL baseline performs very well, at times on par with the best systems existing to date. Given its simplicity and excellent performance, KL should be used as a baseline in future studies.

There are no significant differences between the best systems on R-2 recall. The DPP supervised framework and the unsupervised global optimization approach emerge as the best systems on R-1 and R-2,-4 respectively. Moreover, we show that diverse contents get selected by the summaries from different state-of-the-art systems. This suggests that summary combination might lead to improvements in content selection. Our repository, along with the guidelines on reporting performance, will enable further progress in automatic summarization. The repository is publicly available via this link: https://www.seas.upenn.edu/~nlp/corpora/sumrepo.html .

8. References

Conroy, John M., Goldstein, Jade, Schlesinger, Judith D., and O'leary, Dianne P. (2004). Left-brain/right-brain multi-document summarization. In *Proceedings of the Document Understanding Conference (DUC)*.

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

Conroy, John M, Schlesinger, Judith D, Kubina, Jeff, Rankel, Peter A, and O'Leary, Dianne P. (2011). CLASSY 2011 at TAC: Guided and multi-lingual summaries and evaluation metrics. *Proceedings of the Text Analysis Conference*.

Conroy, John M., Davis, Sashka T., Kubina, Jeff, Liu, Yi-Kai, O'Leary, Dianne P., and Schlesinger, Judith D. (2013). Multilingual summarization: Dimensionality reduction and a step towards optimal term coverage. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 55–63.

Dasgupta, Anirban, Kumar, Ravi, and Ravi, Sujith. (2013). Summarization through submodularity and dispersion. In *Proceedings of ACL*, pages 1014–1022.

Davis, Sashka T., Conroy, John M., and Schlesinger, Judith D. (2012). OCCAMS – An Optimal Combinatorial Covering Algorithm for Multi-document Summarization. In *ICDM Workshops*, pages 454–463.

Erkan, Günes and Radev, Dragomir R. (2004). Lexrank: graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479, December.

Gillick, Dan and Favre, Benoit. (2009). A scalable global model for summarization. In *Proceedings of ACL*

Input d30002t

Greedy-KL Summary: In Washington on Thursday, President Bill Clinton ordered dlrs 30 million in Defense Department equipment and services and dlrs 36 million in food, fuel and other aid be sent to Honduras, Nicaragua, El Salvador and Guatemala. With the storm seemingly anchored off Honduras, officials in Mexico to the north eased emergency measures on the Caribbean coast of the Yucatan Peninsula, where hundreds of people remained in shelters as a precaution Wednesday night. The Honduran president closed schools and public offices on the coast Monday and ordered all air force planes and helicopters to evacuate people from the Islas de la Bahia, a string of small islands off the country's central coast.

DPP Summary: At 0900 GMT Tuesday, Mitch was 95 miles north of Honduras, near the Swan Islands. Hurricane Mitch cut through the Honduran coast like a ripsaw Thursday, **[its devastating winds whirling for a third day]**$_B$ through resort islands and mainland communities. Mexico reported one death from Mitch last Monday. **[Hurricane Mitch killed an estimated 9,000 people throughout Central America]**$_A$ in a disaster of such proportions that relief agencies have been overwhelmed. He urged the more than 1.5 million Hondurans affected by the storm to help with the recovery effort. **[Taiwan said today it will donate dlrs 2.6 million in relief to Honduras, Nicaragua, El Salvador and Guatemala]**$_C$.

ICSISumm Summary: At least 231 people have been confirmed dead in Honduras from former-hurricane Mitch, bringing the storm's death toll in the region to 357. Pope John Paul II appealed for aid Wednesday for the Central American countries stricken by hurricane Mitch and said he feels close to the thousands who are suffering. **[Officials in Central America estimated Tuesday that about 7,000 people have died in the region]**$_A$. **[That meant the Honduran coast had been under hurricane conditions for more than a day]**$_B$. **[Taiwan said today it will donate dlrs 2.6 million in relief to Honduras, Nicaragua, El Salvador and Guatemala]**$_C$.

RegSumm Summary: President Carlos Flores declared Hurricane Mitch had set back Honduras development by 50 years. El Salvador reported 239 dead; Guatemala said 194 of its people had been killed. **[Officials in Central America estimated Tuesday that about 7,000 people have died in the region]**$_A$. **[Hurricane Mitch killed an estimated 9,000 people throughout Central America]**$_A$ in a disaster of such proportions that relief agencies have been overwhelmed. Mexico reported one death from Mitch last Monday. The strongest hurricane to hit Honduras in recent memory was Fifi in 1974, which ravaged Honduras Caribbean coast, killing at least 2,000 people.

Input d30010t

Greedy-KL Summary: **[The radical group Islamic Jihad claimed responsibility Saturday]**$_A$ for the **[suicide bombing of a crowded Jerusalem market]**$_C$ and **[promised more attacks]**$_D$ to **try to block the new peace accord]**$_H$. David Bar-Illan, a top aide to Israeli Prime Minister Benjamin Netanyahu, said Sunday that Israel expects Palestinian leader Yasser Arafat to formally outlaw the military wings of Islamic Jihad and the larger militant group Hamas. **[A Palestinian security official said several Islamic Holy War members were arrested]**$_F$ in the West Bank on Friday night. **[Two people were killed and at least 21 injured]**$_B$ when an explosives-rigged car **[blew up Friday at Jerusalem's Mahane Yehuda market]**$_E$, only meters(yards) from the site of a suicide bombing 16 months ago.

DPP Summary: **[The Israeli Cabinet also announced it will begin to build houses]**$_G$ in the controversial **[Jewish settlement of Har Homa in east Jerusalem]**$_G$. **[The radical group Islamic Jihad claimed responsibility Saturday]**$_A$ for **[the market bombing]**$_C$ and **[vowed more attacks]**$_D$ to **[try to block the new peace accord]**$_H$. Palestinian political leaders said Israel should not use Friday's suicide bombing, **[which killed the two assailants and wounded 21 Israelis]**$_B$, as an excuse to stop the peace process. The militant Palestinian movement Islamic Holy War said Saturday that it **[carried out the suicide bombing in a Jerusalem market on Friday]**$_E$, **[which prompted arrests by the Palestinian Authority overnight]**$_F$.

ICSISumm Summary: A defiant Prime Minister **[Benjamin Netanyahu said Saturday that Israel would continue to build Jewish neighborhoods throughout Jerusalem]**$_G$, including at a controversial site in the traditionally Arab sector of the city. **[The radical group Islamic Jihad claimed responsibility Saturday]**$_A$ for **[the market bombing]**$_C$ and **[vowed more attacks]**$_D$ to **[try to block the new peace accord]**$_H$. Hassan Asfour, a Palestinian peace negotiator, said the Palestinian Authority condemned the attack. An Islamic Jihad official in the Syrian capital of Damascus confirmed that the group's leader Ramadan Abdullah Shallah **[claimed responsibility for Friday's bombing]**$_E$ in an interview with the Paris-based Radio Monte Carlo.

RegSum Summary: **[The radical group Islamic Jihad claimed responsibility Saturday]**$_A$ for **[the suicide bombing of a crowded Jerusalem market]**$_C$ and **[promised more attacks]**$_D$ to **[try to block the new peace accord]**$_H$. The militant Palestinian movement Islamic Holy War said Saturday that **[it carried out the suicide bombing in a Jerusalem market on Friday]**$_E$, **[which prompted arrests by the Palestinian Authority overnight]**$_F$. The Islamic militant group Hamas, which has **[tried to stop the peace agreement]**$_D$, claimed responsibility, police said. **[Two people were killed and 21 others were wounded]**$_B$ in the attack for which **[the Islamic militant group Hamas claimed responsibility]**$_A$.

Table 6: Summaries generated from the Greedy-KL, DPP, ICSISumm and RegSum systems for the input d30002t, d30010t. Contributors of the SCUs which appear in multiple machine summaries are labeled in brackets.

Workshop on Integer Linear Programming for Natural Langauge Processing, pages 10–18.

Gillick, Dan, Favre, Benoit, Hakkani-Tur, Dilek, Bohnet, Berndt, Liu, Yang, and Xie, Shasha. (2009). The ICSI/UTD Summarization System at TAC 2009. In *Proceedings of the Text Understanding Conference*.

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

Hong, Kai and Nenkova, Ani. (2014). Improving the estimation of word importance for news multi-document summarization. In *Proceedings of EACL*, Gothenburg, Sweden, April.

Khuller, Samir, Moss, Anna, and Naor, Joseph. (1999). The budgeted maximum coverage problem. *Inf. Process. Lett.*, 70(1):39–45.

	Greedy-KL	CLASSY 04	CLASSY 11	DPP	ICSISum	OCCAMS_V	RegSum	Submodular
Greedy-KL		(+) 0.6157	(+) 0.3431	**(-) 0.0025**	(-) 0.1992	(-) 0.2754	(-) 0.2362	**(-) 0.0176**
CLASSY 04			(+) 0.7081	**(-) 0.0005**	(-) 0.1145	(-) 0.1296	(-) 0.0800	**(-) 0.0078**
CLASSY 11				**(-) 0.0001**	**(-) 0.0321**	**(-) 0.0057**	**(-) 0.0285**	**(-) 0.0009**
DPP					**(+) 0.0113**	**(+) 0.0049**	**(+) 0.0165**	(+) 0.0719
ICSISum						(-) 0.9714	(-) 0.7906	(-) 0.1346
OCCAMS_V							(-) 0.7779	(-) 0.1925
RegSum								(-) 0.2165
Submodular								

Table 7: P-values for paired two-sided Wilcoxon signed-rank test, on ROUGE-1 recall

	Greedy-KL	CLASSY 04	CLASSY 11	DPP	ICSISum	OCCAMS_V	RegSum	Submodular
Greedy-KL		(-) 0.3884	(-) 0.0836	**(-) 0.0066**	**(-) 0.0002**	**(-) 0.0013**	**(-) 0.0036**	**(-) 0.0277**
CLASSY 04			(-) 0.5211	(-) 0.1331	(-) 0.0710	(-) 0.0572	**(-) 0.0483**	(-) 0.3173
CLASSY 11				(-) 0.1641	(-) 0.0554	(-) 0.1303	(-) 0.1161	(-) 0.6555
DPP					(-) 0.4816	(-) 0.6516	(-) 0.5785	(+) 0.3173
ICSISum						(+) 0.7906	(+) 0.7583	(+) 0.1860
OCCAMS_V							(-) 0.9762	(+) 0.1563
RegSum								(+) 0.1260
Submodular								

Table 8: P-values for paired two-sided Wilcoxon signed-rank test, on ROUGE-2 recall

	Greedy-KL	CLASSY 04	CLASSY 11	DPP	ICSISum	OCCAMS_V	RegSum	Submodular
Greedy-KL		(-) 0.3269	(-) 0.1503	**(-) 0.0373**	**(-) 0.0029**	(-) 0.5531	**(-) 0.0360**	(-) 0.4863
CLASSY 04			(+) 0.6958	(-) 0.6234	(-) 0.1798	(+) 0.7536	(-) 0.4407	(+) 0.7394
CLASSY 11				(-) 0.4687	(-) 0.0806	(+) 0.2672	(-) 0.4396	(+) 0.5183
DPP					(-) 0.2290	(+) 0.1222	(-) 0.9056	**(+) 0.0400**
ICSISum						**(+) 0.0050**	(+) 0.3556	**(+) 0.0095**
OCCAMS_V							(-) 0.1206	(-) 0.4207
RegSum								(+) 0.0660
Submodular								

Table 9: P-values for paired two-sided Wilcoxon signed-rank test, on ROUGE-4 recall

Kulesza, Alex and Taskar, Ben. (2012). Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3).

Lawrence, Page, Sergey, Brin, Motwani, Rajeev, and Winograd, Terry. (1998). The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University.

Lin, Hui and Bilmes, Jeff. (2011). A class of submodular functions for document summarization. In *Proceedings of ACL*, pages 510–520.

Lin, Hui and Bilmes, Jeff. (2012). Learning mixtures of submodular shells with application to document summarization. In *Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, USA, July. AUAI.

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

Lin, Chin-Yew. (2004). ROUGE: a package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Louis, Annie and Nenkova, Ani. (2013). Automatically assessing machine summary content without a gold standard. *Comput. Linguist.*, 39(2):267–300, June.

Manna, Sukanya, Gao, Byron J., and Coke, Reed. (2012). A subjective logic framework for multi-document summarization. In *Proceedings of COLING 2012: Posters*, pages 797–808.

Mihalcea, Rada and Tarau, Paul. (2004). Textrank: Bringing order into text. In *Proceedings of EMNLP*, pages 404–411.

Morita, Hajime, Sasano, Ryohei, Takamura, Hiroya, and Okumura, Manabu. (2013). Subtree extractive summarization via submodular maximization. In *Proceedings of ACL*, pages 1023–1032.

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

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

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

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

Owczarzak, Karolina, Conroy, John M., Dang, Hoa Trang, and Nenkova, Ani. (2012). An assessment of the accuracy of automatic evaluation in summarization. In *NAACL-HLT 2012: Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9.

Radev, D., Allison, T., Goldensohn, Blair S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E., Hakim, A., Lam, W., and Liu, D. (2004a). MEAD: a platform for multidocument multilingual text summarization. In *Proceedings of LREC*, pages 1–4.

Radev, Dragomir R., Jing, Hongyan, Stys, Malgorzata, and Tam, Daniel. (2004b). Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938, November.

Rankel, Peter, Conroy, John, Slud, Eric, and O'Leary, Dianne. (2011). Ranking human and machine summarization systems. In *Proceedings of EMNLP*, pages 467–473.

Rankel, Peter A., Conroy, John M., Dang, Hoa Trang, and Nenkova, Ani. (2013). A decade of automatic content evaluation of news summaries: Reassessing the state of the art. In *Proceedings of ACL (Volume 2: Short Papers)*, pages 131–136.

Sandhaus, Evan. (2008). The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia, PA*.

Sipos, Ruben, Shivaswamy, Pannaga, and Joachims, Thorsten. (2012). Large-margin learning of submodular summarization models. In *Proceedings of EACL*, pages 224–233.

Wan, Xiaojun and Yang, Jianwu. (2008). Multi-document summarization using cluster-based link analysis. In *Proceedings of SIGIR*, pages 299–306.