Discrepancy Between Automatic and Manual Evaluation of Summaries

Shamima Mithun, Leila Kosseim, and Prasad Perera
Concordia University
Department of Computer Science and Software Engineering
Montreal, Quebec, Canada
{s_mithun, kosseim, p_perer}@encs.concordia.ca

Abstract

Today, automatic evaluation metrics such as ROUGE have become the de-facto mode of evaluating an automatic summarization system. However, based on the DUC and the TAC evaluation results, (Conroy and Schlesinger, 2008; Dang and Owczarzak, 2008) showed that the performance gap between human-generated summaries and system-generated summaries is clearly visible in manual evaluations but is often not reflected in automated evaluations using ROUGE scores. In this paper, we present our own experiments in comparing the results of manual evaluations versus automatic evaluations using our own text summarizer: BlogSum. We have evaluated BlogSum-generated summary content using ROUGE and compared the results with the original candidate list (OList). The t-test results showed that there is no significant difference between BlogSum-generated summaries and OList summaries. However, two manual evaluations for content using two different datasets show that BlogSum performed significantly better than OList. A manual evaluation of summary coherence also shows that BlogSum performs significantly better than OList. These results agree with previous work and show the need for a better automated summary evaluation metric rather than the standard ROUGE metric.

1 Introduction

Today, any NLP task must be accompanied by a well-accepted evaluation scheme. This is why, for the last 15 years, to evaluate automated summarization systems, sets of evaluation data (corpora, topics, ...) and baselines have been established in text summarization competitions such as TREC[1], DUC[2], and TAC[3]. Although evaluation is essential to verify the quality of a summary or to compare different summarization approaches, the evaluation criteria used are by no means universally accepted (Das and Martins, 2007). Summary evaluation is a difficult task because no ideal summary is available for a set of input documents. In addition, it is also difficult to compare different summaries and establish a baseline because of the absence of standard human or automatic summary evaluation metrics. On the other hand, manual evaluation is very expensive. According to (Lin, 2004), large scale manual evaluations of all participants' summaries in the DUC 2003 conference would require over 3000 hours of human efforts to evaluate summary content and linguistic qualities.

The goal of this paper is to show that the literature and our own work empirically point out the need for a better automated summary evaluation metric rather than the standard ROUGE metric[4] (Lin, 2004).

2 Current Evaluation Schemes

The available summary evaluation techniques can be divided into two categories: manual and automatic. To do a manual evaluation, human experts assess different qualities of the system generated summaries. On the other hand, for an automatic eval-

[1]Text REtrieval Conference: http://trec.nist.gov
[2]Document Understanding Conference: http://duc.nist.gov
[3]Text Analysis Conference: http://www.nist.gov/tac
[4]http://berouge.com/default.aspx

Proceedings of the Workshop on Evaluation Metrics and System Comparison for Automatic Summarization, pages 44–52,
Montréal, Canada, June 3-8, 2012.

uation, tools are used to compare the system generated summaries with human generated gold standard summaries or reference summaries. Although they are faster to perform and result in consistent evaluations, automatic evaluations can only address superficial concepts such as n-grams matching, because many required qualities such as coherence and grammaticality cannot be measured automatically. As a result, human judges are often called for to evaluate or cross check the quality of the summaries, but in many cases human judges have different opinions. Hence inter-annotator agreement is often computed as well.

The quality of a summary is assessed mostly on its content and linguistic quality (Louis and Nenkova, 2008). Content evaluation of a query-based summary is performed based on the relevance with the topic and the question and the inclusion of important contents from the input documents. The linguistic quality of a summary is evaluated manually based on how it structures and presents the contents. Mainly, subjective evaluation is done to assess the linguistic quality of an automatically generated summary. Grammaticality, non-redundancy, referential clarity, focus, structure and coherence are commonly used factors considered to evaluate the linguistic quality. A study by (Das and Martins, 2007) shows that evaluating the content of a summary is more difficult compared to evaluating its linguistic quality.

There exist different measures to evaluate an output summary. The most commonly used metrics are *recall*, *precision*, *F-measure*, *Pyramid score*, and *ROUGE/BE*.

Automatic versus Manual Evaluation

Based on an analysis of the 2005-2007 DUC data, (Conroy and Schlesinger, 2008) showed that the ROUGE evaluation and a human evaluation can significantly vary due to the fact that ROUGE ignores linguistic quality of summaries, which has a huge influence in human evaluation. (Dang and Owczarzak, 2008) also pointed out that automatic evaluation is rather different than the one based on manual assessment. They explained this the following way: "automatic metrics, based on string matching, are unable to appreciate a summary that uses different phrases than the reference text, even if such a summary is perfectly fine by human standards".

To evaluate both opinionated and news article based summarization approaches, previously mentioned evaluation metrics such as ROUGE or Pyramid are used. Shared evaluation tasks such as DUC and TAC competitions also use these methods to evaluate participants' summary. Table 1 shows the evaluation results of automatic systems' average performance at the TAC 2008 to 2010 conferences using the pyramid score (Pyr.) and responsiveness (Resp.). In this evaluation, the pyramid score was used to calculate the content relevance and the responsiveness of a summary was used to judge the overall quality or usefulness of the summary, considering both the information content and linguistic quality. These two criteria were evaluated manually. The pyramid score was calculated out of 1 and the responsiveness measures were calculated on a scale of 1 to 5 (1, being the worst). However, in 2009, responsiveness was calculated on a scale of 10. Table 1 also shows a comparison between automatic systems and human participants (model). In Table 1, the first 3 rows show the evaluation results of the TAC Update Summarization (Upd.) initial summary generation task (which were generated for news articles) and the last row shows the evaluation results of the TAC 2008 Opinion Summarization track (Opi.) where summaries were generated from blogs. From Table 1, we can see that in both criteria, automatic systems are weaker than humans. (Note that in the table, Unk. refers to unknown.)

Table 1: Human and Automatic System Performance at Various TAC Competitions

	Model (Human)		**Automatic**	
	Pyr.	**Resp.**	**Pyr.**	**Resp.**
2010 Upd.	0.78	4.76	0.30	2.56
2009 Upd.	0.68	8.83	0.26	4.14
2008 Upd.	0.66	4.62	0.26	2.32
2008 Opi.	0.44	Unk.	0.10	1.31

Interestingly, in an automatic evaluation, often, not only is there no significant gap between models and systems, but in many cases, automatic systems scored higher than some human models.

Table 2 shows the performance of human (H.) and automated systems (S.) (participants) using automated and manual evaluation in the TAC 2008 up-

Table 2: Automated vs. Manual Evaluation at TAC 2008

	Automated		Manual		
	R-2	**R-SU4**	**Pyr.**	**Ling.**	**Resp.**
H. Mean	0.12	0.15	0.66	4.79	4.62
S. Mean	0.08	0.12	0.26	2.33	2.32
H. Best	0.13	0.17	0.85	4.91	4.79
S. Best	0.11	0.14	0.36	3.25	2.29

date summarization track. In the table, R-2 and R-SU4 refer to ROUGE-2 and ROUGE-SU4 and Pyr., Ling., and Resp. refer to Pyramid, linguistic, and responsiveness, respectively. A *t*-test of statistical significance applied to the data in Table 2 shows that there is no significant difference between human and participants in automated evaluation but that there is a significant performance difference between them in the manual evaluation.

These findings indicate that ROUGE is not the most effective tool to evaluate summaries. Our own experiments described below arrive at the same conclusion.

3 BlogSum

We have designed an extractive query-based summrizer called BlogSum. In BlogSum, we have developed our own sentence extractor to retrieve the initial list of candidate sentences (we called it OList) based on question similarity, topic similarity, and subjectivity scores. Given a set of initial candidate sentences, BlogSum generates summaries using discourse relations within a schema-based framework. Details of BlogSum is outside the scope of this paper. For details, please see (Mithun and Kosseim, 2011).

4 Evaluation of BlogSum

BlogSum-generated summaries have been evaluated for content and linguistic quality, specifically discourse coherence. The evaluation of the content was done both automatically and manually and the evaluation of the coherence was done manually. Our evaluation results also reflect the discrepancy between automatic and manual evaluation schemes of summaries described above.

In our evaluation, BlogSum-generated summaries were compared with the original candidate list generated by our approach without the discourse reordering (OList). However, we have validated our original candidate list with a publicly available sentence ranker. Specifically, we have conducted an experiment to verify whether MEAD-generated summaries (Radev et al., 2004), a widely used publicly available summarizer[5], were better than our candidate list (OList). In this evaluation, we have generated summaries using MEAD with centroid, query title, and query narrative features. In MEAD, query title and query narrative features are implemented using cosine similarity based on the *tf-idf* value. In this evaluation, we used the TAC 2008 opinion summarization dataset (described later in this section) and summaries were evaluated using the ROUGE-2 and ROUGE-SU4 scores. Table 3 shows the results of the automatic evaluation using ROUGE based on summary content.

Table 3: Automatic Evaluation of MEAD based on Summary Content on TAC 2008

System	**R-2 (F)**	**R-SU4 (F)**
MEAD	0.0407	0.0642
Average	0.0690	0.0860
OList	0.1020	0.1070

Table 3 shows that MEAD-generated summaries achieved weaker ROUGE scores compared to that of our candidate list (OList). The table also shows that MEAD performs weaker than the average performance of the participants of TAC 2008 (Average). We suspect that these poor results are due to several reasons. First, in MEAD, we cannot use opinionated terms or polarity information as a sentence selection feature. On the other hand, most of the summarizers, which deal with opinionated texts, use opinionated terms and polarity information for this purpose. In addition, in this experiment, for some of the TAC 2008 questions, MEAD was unable to create any summary. This evaluation results prompted us to develop our own candidate sentence selector.

[5]MEAD: http://www.summarization.com/mead

4.1 Evaluation of Content

4.1.1 Automatic Evaluation of Content

First, we have automatically evaluated the summaries generated by our approach for content. As a baseline, we used the original ranked list of candidate sentences (OList), and compared them to the final summaries (BlogSum). We have used the data from the TAC 2008 opinion summarization track for the evaluation.

The dataset consists of 50 questions on 28 topics; on each topic one or two questions are asked and 9 to 39 relevant documents are given. For each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words. This length was chosen cause in the DUC conference from 2005 to 2007, in the main summarization task, the summary length was 250 words. In addition, (Conroy and Schlesinger, 2008) also created summaries of length 250 words in their participation in the TAC 2008 opinion summarization task and performed well. (Conroy and Schlesinger, 2008) also pointed out that if the summaries were too long this adversely affected their scores. Moreover, according to the same authors shorter summaries are easier to read. Based on these observations, we have restricted the maximum summary length to 250 words. However, in the TAC 2008 opinion summarization track, the allowable summary length is very long (the number of non-whitespace characters in the summary must not exceed 7000 times the number of questions for the target of the summary). In this experiment, we used the ROUGE metric using answer nuggets (provided by TAC), which had been created to evaluate participants' summaries at TAC, as gold standard summaries. F-scores are calculated for BlogSum and OList using ROUGE-2 and ROUGE-SU4. In this experiment, ROUGE scores are also calculated for all 36 submissions in the TAC 2008 opinion summarization track.

The evaluation results are shown in Table 4. Note that in the table *Rank* refers to the rank of the system compared to the other 36 systems.

Table 4 shows that BlogSum achieved a better F-Measure (F) for ROUGE-2 (R-2) and ROUGE-SU4 (R-SU4) compared to OList. From the results, we can see that BlogSum gained 18% and 16% in F-Measure over OList using ROUGE-2 and ROUGE-SU4, respectively.

Table 4: Automatic Evaluation of BlogSum based on Summary Content on TAC 2008

System	R-2 (F)	R-SU4 (F)	Rank
Best	0.130	0.139	1
BlogSum	0.125	0.128	3
OList	0.102	0.107	10
Average	0.069	0.086	N/A

Compared to the other systems that participated to the TAC 2008 opinion summarization track, BlogSum performed very competitively; it ranked third and its F-Measure score difference from the best system is very small. Both BlogSum and OList performed better than the average systems.

However, a further analysis of the results of Table 4 shows that there is no significant difference between BlogSum-generated summaries and OList summaries using the t-test with a *p-value* of 0.228 and 0.464 for ROUGE-2 and ROUGE-SU4, respectively. This is inline with (Conroy and Schlesinger, 2008; Dang and Owczarzak, 2008) who showed that the performance gap between human-generated summaries and system-generated summaries at DUC and TAC is clearly visible in a manual evaluation, but is often not reflected in automated evaluations using ROUGE scores. Based on these findings, we suspected that there might be a performance difference between BlogSum-generated summaries and OList which is not reflected in ROUGE scores. To verify our suspicion, we have conducted manual evaluations for content.

4.1.2 Manual Evaluation of Content using the Blog Dataset

We have conducted two manual evaluations using two different datasets to better quantify BlogSum-generated summary content.

Corpora and Experimental Design

In the first evaluation, we have again used the TAC 2008 opinion summarization track data. For each question, one summary was generated by OList and one by BlogSum and the maximum summary length was again restricted to 250 words. To evaluate

content, 3 participants manually rated 50 summaries from OList and 50 summaries from BlogSum using a blind evaluation. These summaries were rated on a likert scale of 1 to 5 where 1 refers to "very poor" and 5 refers to "very good". Evaluators rated each summary with respect to the question for which it was generated and against the reference summary. In this experiment, we have used the answer nuggets provided by TAC as the reference summary, which had been created to evaluate participants' summaries at TAC. Annotators were asked to evaluate summaries based on their content without considering their linguistic qualities.

Results

In this evaluation, we have calculated the average scores of all 3 annotators' ratings to a particular question to compute the score of BlogSum for a particular question. Table 5 shows the performance comparison between BlogSum and OList. The results show that 58% of the time BlogSum summaries were rated better than OList summaries which implies that 58% of the time, our approach has improved the question relevance compared to that of the original candidate list (OList).

Table 5: Comparison of OList and BlogSum based on the Manual Evaluation of Summary Content on TAC 2008

Comparison	%
BlogSum Score > OList Score	58%
BlogSum Score = OList Score	30%
BlogSum Score < OList Score	12%

Table 6 shows the performance of BlogSum versus OList on each likert scale; where Δ shows the difference in performance. Table 6 demonstrates that 52% of the times, BlogSum summaries were rated as "very good" or "good", 26% of the times they were rated as "barely acceptable" and 22% of the times they were rated as "poor" or "very poor". From Table 6, we can also see that BlogSum outperformed OList in the scale of "very good" and "good" by 8% and 22%, respectively; and improved the performance in "barely acceptable", "poor", and "very poor" categories by 12%, 8%, and 10%, respectively.

In this evaluation, we have also calculated whether there is any performance gap between BlogSum and OList. The *t*-test results show that in a two-tailed test, BlogSum performed significantly better than OList with a *p*-value of 0.00281.

Table 6: Manual Evaluation of BlogSum and OList based on Summary Content on TAC 2008

Category	OList	BlogSum	Δ
Very Good	6%	14%	8%
Good	16%	38%	22%
Barely Acceptable	38%	26%	-12%
Poor	26%	18%	-8%
Very Poor	14%	4%	-10%

Whenever human performance is computed by more than one person, it is important to compute inter-annotator agreement. This ensures that the agreement between annotators did not simply occur by chance. In this experiment, we have also calculated the inter-annotator agreement using Cohen's kappa coefficient to verify the annotation subjectivity. We have found that the average pair-wise inter-annotator agreement is moderate according to (Landis and Koch, 1977) with the kappa-value of 0.58.

4.1.3 Manual Evaluation of Content using the Review Dataset

We have conducted a second evaluation using the OpinRank dataset[6] and (Jindal and Liu, 2008)'s dataset to evaluate BlogSum-generated summary content.

Corpora and Experimental Design

In this second evaluation, we have used a subset of the OpinRank dataset and (Jindal and Liu, 2008)'s dataset. The OpinRank dataset contains reviews on cars and hotels collected from Tripadvisor (about 259,000 reviews) and Edmunds (about 42,230 reviews). The OpinRank dataset contains 42,230 reviews on cars for different model-years and 259,000 reviews on different hotels in 10 different cities. For this dataset, we created a total of 21 questions including 12 reason questions and 9 suggestions. For each question, 1500 to 2500 reviews were provided

[6]OpinRank Dataset: http://kavita-ganesan.com/entity-ranking-data

as input documents to create the summary.

(Jindal and Liu, 2008)'s dataset consists of 905 comparison and 4985 non-comparison sentences. Four human annotators labeled these data manually. This dataset consists of reviews, forum, and news articles on different topics from different sources. We have created 9 comparison questions for this dataset. For each question, 700 to 1900 reviews were provided as input documents to create the summary.

For each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words again. To evaluate question relevance, 3 participants manually rated 30 summaries from OList and 30 summaries from BlogSum using a blind evaluation. These summaries were again rated on a likert scale of 1 to 5. Evaluators rated each summary with respect to the question for which it was generated.

Results

Table 7 shows the performance comparison between BlogSum and OList. The results show that 67% of the time BlogSum summaries were rated better than OList summaries. The table also shows that 30% of the time both approaches performed equally well and 3% of the time BlogSum was weaker than OList.

Table 7: Comparison of OList and BlogSum based on the Manual Evaluation of Summary Content on the Review Dataset

Comparison	%
BlogSum Score > OList Score	67%
BlogSum Score = OList Score	30%
BlogSum Score < OList Score	3%

Table 8 demonstrates that 44% of the time BlogSum summaries were rated as "very good", 33% of the time rated as "good", 13% of the time they were rated as "barely acceptable" and 10% of the time they were rated as "poor" or "very poor". From Table 8, we can also see that BlogSum outperformed OList in the scale of "very good" by 34% and improved the performance in "poor" and "very poor" categories by 23% and 10%, respectively.

In this evaluation, we have also calculated whether there is any performance gap between BlogSum and OList. The *t*-test results show that in a two-tailed test, BlogSum performed significantly very better than OList with a *p*-value of 0.00236. In addition, the average pair-wise inter-annotator agreement is substantial according to (Landis and Koch, 1977) with the kappa-value of 0.77.

Table 8: Manual Evaluation of BlogSum and OList based on Summary Content on the Review Dataset

Category	OList	BlogSum	Δ
Very Good	10%	44%	34%
Good	37%	33%	-4%
Barely Acceptable	10%	13%	3%
Poor	23%	0%	-23%
Very Poor	20%	10%	-10%

4.1.4 Analysis

In both manual evaluation for content, BlogSum performed significantly better than OList. We can see that even though there was not any significant performance gap between BlogSum and OList-generated summaries in the automatic evaluation of Section 4.1.1, both manual evaluations show that BlogSum and OList-generated summaries significantly vary at the content level. For content, our results support (Conroy and Schlesinger, 2008; Dang and Owczarzak, 2008)'s findings and points out for a better automated summary evaluation tool.

4.2 Evaluation of Linguistic Quality

Our next experiments were geared at evaluating the linguistic quality of our summaries.

4.2.1 Automatic Evaluation of Linguistic Quality

To test the linguistic qualities, we did not use an automatic evaluation because (Blair-Goldensohn and McKeown, 2006) found that the ordering of content within the summaries is an aspect which is not evaluated by ROUGE. Moreover, in the TAC 2008 opinion summarization track, on each topic, answer snippets were provided which had been used as summarization content units (SCUs) in pyramid evaluation to evaluate TAC 2008 participants summaries but no complete summaries is provided to which we can compare BlogSum-generated summaries for co-

herence. As a result, we only performed two manual evaluations using two different datasets again to see whether BlogSum performs significantly better than OList for linguistic qualities too. The positive results of the next experiments will ensure that BlogSum-generated summaries are really significantly better than OList summaries.

4.2.2 Manual Evaluation of Discourse Coherence using the Blog Dataset

In this evaluation, we have again used the TAC 2008 opinion summarization track data. For each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words again. Four participants manually rated 50 summaries from OList and 50 summaries from BlogSum for coherence. These summaries were again rated on a likert scale of 1 to 5.

Results

To compute the score of BlogSum for a particular question, we calculated the average scores of all annotators' ratings to that question. Table 9 shows the performance comparison between BlogSum and OList. We can see that 52% of the time BlogSum summaries were rated better than OList summaries; 30% of the time both performed equally well; and 18% of the time BlogSum was weaker than OList. This means that 52% of the time, our approach has improved the coherence compared to that of the original candidate list (OList).

Table 9: Comparison of OList and BlogSum based on the Manual Evaluation of Discourse Coherence on TAC 2008

Comparison	%
BlogSum Score > OList Score	52%
BlogSum Score = OList Score	30%
BlogSum Score < OList Score	18%

From Table 10, we can see that BlogSum outperformed OList in the scale of "very good" and "good" by 16% and 8%, respectively; and improved the performance in "barely acceptable" and "poor" categories by 12% and 14%, respectively.

The *t*-test results show that in a two-tailed test, BlogSum performed significantly better than OList with a *p*-value of 0.0223. In addition, the average pair-wise inter-annotator agreement is substantial according to with the kappa-value of 0.76.

Table 10: Manual Evaluation of BlogSum and OList based on Discourse Coherence on TAC 2008

Category	OList	BlogSum	Δ
Very Good	8%	24%	16%
Good	22%	30%	8%
Barely Acceptable	36%	24%	-12%
Poor	22%	8%	-14%
Very Poor	12%	14%	2%

4.2.3 Manual Evaluation of Discourse Coherence using the Review Dataset

In this evaluation, we have again used the OpinRank dataset and (Jindal and Liu, 2008)'s dataset to conduct the second evaluation of content. In this evaluation, for each question, one summary was generated by OList and one by BlogSum and the maximum summary length was restricted to 250 words. Three participants manually rated 30 summaries from OList and 30 summaries from BlogSum for coherence.

Results

To compute the score of BlogSum for a particular question, we calculated the average scores of all annotators' ratings to that question. Table 11 shows the performance comparison between BlogSum and OList. We can see that 57% of the time BlogSum summaries were rated better than OList summaries; 20% of the time both performed equally well; and 23% of the time BlogSum was weaker than OList.

Table 11: Comparison of OList and BlogSum based on the Manual Evaluation of Discourse Coherence on the Review Dataset

Comparison	%
BlogSum Score > OList Score	57%
BlogSum Score = OList Score	20%
BlogSum Score < OList Score	23%

Table 12: Manual Evaluation of BlogSum and OList based on Discourse Coherence on the Review Dataset

Category	OList	BlogSum	Δ
Very Good	13%	23%	10%
Good	27%	43%	16%
Barely Acceptable	27%	17%	-10%
Poor	10%	10%	0%
Very Poor	23%	7%	-16%

From Table 12, we can see that BlogSum outperformed OList in the scale of "very good" and "good" by 10% and 16%, respectively; and improved the performance in "barely acceptable" and "very poor" categories by 10% and 16%, respectively.

We have also evaluated if the difference in performance between BlogSum and OList was statistically significant. The *t*-test results show that in a two-tailed test, BlogSum performed significantly better than OList with a *p*-value of 0.0371.

In this experiment, we also calculated the inter-annotator agreement using Cohen's kappa coefficient. We have found that the average pair-wise inter-annotator agreement is substantial according to (Landis and Koch, 1977) with the kappa-value of 0.74.

The results of both manual evaluations of discourse coherence also show that BlogSum performs significantly better than OList.

5 Conclusion

Based on the DUC and TAC evaluation results, (Conroy and Schlesinger, 2008; Dang and Owczarzak, 2008) showed that the performance gap between human-generated summaries and system-generated summaries, which is clearly visible in the manual evaluation, is often not reflected in automated evaluations using ROUGE scores. In our content evaluation, we have used the automated measure ROUGE (ROUGE-2 & ROUGE-SU4) and the t-test results showed that there was no significant difference between BlogSum-generated summaries and OList summaries with a *p-value* of 0.228 and 0.464 for ROUGE-2 and ROUGE-SU4, respectively. We suspected that there might be a performance difference between BlogSum-generated summaries and OList which is not reflected in ROUGE scores. To verify our suspicion, we have conducted two manual evaluations for content using two different datasets. The *t*-test results for both datasets show that in a two-tailed test, BlogSum performed significantly better than OList with a *p*-value of 0.00281 and 0.00236. Manual evaluations of coherence also show that BlogSum performs significantly better than OList. Even though there was no significant performance gap between BlogSum and OList-generated summaries in the automatic evaluation, the manual evaluation results clearly show that BlogSum-generated summaries are better than OList significantly. Our results supports (Conroy and Schlesinger, 2008; Dang and Owczarzak, 2008)'s findings and points out for a better automated summary evaluation tool.

Acknowledgement

The authors would like to thank the anonymous referees for their valuable comments on a previous version of the paper.
This work was financially supported by NSERC.

References

Annie Louis and Ani Nenkova. 2008. *Automatic Summary Evaluation without Human Models.* Proceedings of the First Text Analysis Conference (TAC 2008), Gaithersburg, Maryland (USA), November.

Chin-Y. Lin. 2004. *ROUGE: A Package for Automatic Evaluation of Summaries.* Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pages 74–81, Barcelona, Spain, July.

Dipanjan Das and Andre F. T. Martins. 2007. *A Survey on Automatic Text Summarization.* Available from: http://www.cs.cmu.edu/∼ nasmith/LS2/ das-martins.07.pdf, Literature Survey for the Language and Statistics II course at Carnegie Mellon University.

Dragomir Radev et al. 2004. *MEAD -A Platform for Multidocument Multilingual Text Summarization.* Proceedings of the the 4th International Conference on Language Resources and Evaluation, pages 1–4, Lisbon, Portugal.

Hoa T. Dang and Karolina Owczarzak. 2008. *Overview of the TAC 2008 Update Summarization Task.* Proceedings of the Text Analysis Conference, Gaithersburg, Maryland (USA), November.

John M. Conroy and Judith D. Schlesinger. 2008. *CLASSY and TAC 2008 Metrics.* Proceedings of the

Text Analysis Conference, Gaithersburg, Maryland (USA), November.

John M. Conroy and Hoa T. Dang. 2008. *Mind the Gap: Dangers of Divorcing Evaluations of Summary Content from Linguistic Quality.* Proceedings of the the 22nd International Conference on Computational Linguistics Coling, pages 145–152, Manchester, UK.

Nitin Jindal and Bing Liu. 2006. *Identifying Comparative Sentences in Text Documents.* SIGIR'06 Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 244–251, Seattle, Washington, USA, August.

Richard J. Landis and Gary G. Koch. 1977. *A One-way Components of Variance Model for Categorical Data. Journal of Biometrics*, 33(1):671–679.

Sasha Blair-Goldensohn and Kathleen McKeown. 2006. *Integrating Rhetorical-Semantic Relation Models for Query-Focused Summarization.* Proceedings of the Document Understanding Conference (DUC) Workshop at NAACL-HLT 2006, New York, USA, June.

Shamima Mithun and Leila Kosseim. 2011. *Discourse Structures to Reduce Discourse Incoherence in Blog Summarization.* Proceedings of Recent Advances in Natural Language Processing, pages 479–486, Hissar, Bulgaria, September.