Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Ranking Summaries for Informativeness and Coherence without Reference Summaries

Abhishek Singh and Wei Jin*

Chegg Inc., Santa Clara, CA, USA
*Department of Computer Science, North Dakota State University, Fargo, ND, USA
abhishek@chegg.com, wei.jin@ndsu.edu

Abstract

There are numerous applications of automatic summarization systems currently and evaluating the quality of the summary is an important task. Current summary evaluation methods are limited in their scope since they rely on a reference summary, i.e., a human written summary. In this paper, we present a new summary evaluation technique without the use of reference summaries. The framework consists of two sequential steps: feature extraction and rank learning and generation. The former extracts effective features reflecting generic aspects, coherence, topical relevance, and informativeness of summaries and the latter uses features to train a learning model that provides the capability of generating a pair wise ranking for input summaries automatically. Our proposed framework is evaluated on the DUC multi-document summarization dataset and results indicate that this is a promising direction for automatic evaluation of the summaries without the use of a reference summary.

Introduction

Automatic multi-document summarization and evaluation has drawn much attention in recent years. In the communities of natural language processing and information retrieval, a series of workshops and conferences on automatic text summarization (e.g., NTCIR, DUC, and TAC), special topic sessions in ACL, NACCL-HLT, CIKM, COLING, and SIGIR have advanced the summarization techniques and produced experimental online systems. Shared tasks organized by NTCIR and DUC/TAC have provided evaluation of system-generated summaries. These evaluations involve human judgments of the system summaries against human written target summaries; these judgments comprise responsiveness, coherence, and linguistic quality. Due to the cost of these manual evaluations, the field has been looking for approximate automatic evaluation measures, especially to enable system development. Current automatic evaluation techniques are characterized by the same need for a reference summary. The most popular evaluation technique involves automatically measuring the content overlap between a human-generated summary and a system-generated summary. The degree of content overlap between the two is measured by some variation of n-gram overlaps such as the Bleu system (Papineni et al., 2002) and the more widely used Rouge system (Lin and Hovy, 2003; Lin 2004). However, in many real applications, reference summaries may not be available, in which case current evaluation approaches cannot be effectively employed. Our goal in this research is to develop a standalone evaluation model, that is, one that does not require a topic-specific reference summary.

Another contribution is to propose a new evaluation scheme that incorporates additional criteria such as coherence (Barzilay and Lapata, 2005), topical relevance (Rahimi et al., 2015), and informativeness (i.e., intrinsic importance of content), in addition to content overlap. The coherence criterion is introduced based on the fact that many system-generated multi-document summaries, while reflecting appropriate content and scoring reasonably well in automatic evaluation measures, score poorly on manual judgments such as coherence. The topical relevance criterion is proposed based on the observation that the document collection can be represented as a mixture of underlying topics with a probability distribution representing the importance of each topic for the collection. Therefore the best terms included in the summary should be those that relate one document to other topically similar documents. This generally holds true, as a good summary should consist of multiple topics reflected in the document set.

The proposed evaluation framework involves two sequential steps: feature extraction and rank learning and generation. After collecting features described above on a corpus of summaries, a learning model is trained based on

manually judged pairs of summaries in the training set, and then our best summary evaluation model is applied to the test data to provide a pair wise ranking for input summaries automatically. The pair wise rankings are finally merged into a single ranking that can then be compared with rankings produced from other alternative evaluation methods.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 describes the process of feature selection and extraction. Section 4 presents the rank generation and evaluation strategy in more detail. Section 5 discusses evaluation datasets and experimental results. Finally sections 6 presents the conclusion.

Related Work

In addition to the relevant literature introduced in Section 1, Barzilay and Lapata (2005) presented a model for evaluating summaries based on coherence. To compute coherence, they used patterns of local entity transitions. A *local entity transition* is a sequence {S, O, X, –}^{n} that represents entity occurrences and their syntactic roles in *n* adjacent sentences. While our method takes coherence and informativeness into account, Barzilay and Lapata focused exclusively on coherence, requiring manually judged coherence values for each summary.

Higgins and Burstein (2006) presented a method to improve on Latent Semantic Analysis (LSA)-based scoring of college essays by using random indexing to assess textual coherence. They claimed that this provided similar results compared to more linguistically oriented techniques while reducing the computational cost. Coh-Metrix (Graesser et al., 2004) was presented to produce various indices that can be combined in several ways in order to assess cohesion of the text, as well as the coherence of the underlying mental representation. Cohesion here is well defined as characteristics of surface level text that helps readers connect ideas. Recently Rahimi et al. (2015) designed a task-dependent model that aligns with the scoring rubric and makes use of the source material. It also incorporates coherence of topics as a criterion in automatic response-to-text assessment of the organization of writing.

Jagarlamudi et al. (2006) presented a different method to score sentences (and hence summaries) using a combination of both query-dependent and query-independent features. Specifically, a method to compute query independent sentence importance is designed based on learning a language model from representative sentences obtained from the Web. By incorporating this score along with query-dependent scoring, their system performed well in DUC competition. They did not use any coherence features, however. In our research, an evaluation framework that incorporates query-dependent and query independent features, as well as coherence and informativeness features, has been provided. The use of coherence features helps us quantify if the sentences in the summary are coherent.

Incorporating topical relevance scores into summary evaluation is another focus of our proposed model. The importance of usage of the topic modeling technique for multiple document summarizations has received much attention. Lau et al. (2014) examined various methodologies that estimate the semantic interpretability of topics at two levels: the model level and the topic level. Haghighi and Vanderwende (2009) described a hierarchical LDA-based model to represent content specificity. The model is constructed based on a hierarchy of topic vocabulary distributions and yields the state-of-the-art ROUGE performance. Chen et al. (2009) presented a novel Bayesian topic model for learning discourse-level document structure which has been proven to outperform other state-of-the-art models in their comparative study. More recently, Yang et al. (2015) proposed a novel contextual topic model for multi-document summarization. The model incorporates the concepts of n-grams into hierarchically latent topics to capture the word dependencies that appear in the local context of a word. In our framework, topic level information is also considered as an important feature, which is inferred from the corpus first and then incorporated into the framework as a distinct measure to evaluate topical relevance of generated summaries to a document collection.

Feature Selection and Extraction

This section describes the various features that are used in the evaluation model. The proposed features are aimed to capture informativeness, coherence, and topical relevance of a summary, in the absence of a reference summary. For the model development we used the DUC (Document Understanding Conference, now a summarization track in the Text Analysis Conference (TAC)) multi-document summarization dataset that consists of 50 document sets or queries. Each query has around 25-50 relevant documents.

Informativeness Features

These features are computed in an attempt to assess the informativeness of the content in a summary. Some of these features are computed based on analysis of a corpus of relevant documents, in this case, the DUC document collection relevant to a query.

Inverse Document Frequency (IDF): the average of inverse document frequency (IDF) values, computed from the corpus, for all terms in a summary. Summaries containing specialized terms are weighted higher.

$$AVG - IDF(S) = \frac{1}{N}\sum_{w \in S} IDF(w)$$

Where N is the number of terms in the summary S and $IDF(w)$ is the IDF value for term w.

Concreteness: Concrete words invite the mental capacity to form images (e.g., glittering diamond), whereas abstract language has relatively less capacity to do so (e.g., conceptual thought). The mean concreteness value of all the words in the summary that have a match in the MRC database was used. The MRC Psycholinguistics Database contains 150,837 words and provides information of up to 26 different linguistic properties of these words, including concreteness. High numbers lean toward concrete and low numbers to abstract.

Pseudo Summary Similarity: In an attempt to automatically simulate a reference summary, the introductory paragraph of each document in the corpus is extracted and synthesized into a single document representing the pseudo summary. It should be noted that DUC produces an automatic baseline summary by taking the first 250 words of the most recent document in the collection. Since this information is not always available, we choose to sample all documents. Presumably, the first paragraph, as the introduction of a well-written document, generally tends to summarize the document. We apply a *Bag of Words* model to all the words in a pseudo summary and then compute the cosine similarity between the summary being evaluated and this pseudo summary as a feature.

$$Sim(S) = \cos(\ pseudoSum\ , S)$$

SumBasic: Based on the SUMBASIC algorithm proposed by Nenkova and Vanderwende (2005), we compute the average weight of all the sentences in the summary. The design of SUMBASIC is motivated by the observation that the relative frequency of a non-stop word is a good indicator of a summary word.

$$Avg - Score(S) = \frac{1}{N}\sum_{e \in S}\sum_{w \in e} P(w) \times IDF(w)$$

Where N is the number of sentences in the summary and $P(w)$ is the unigram probability obtained from the corpus.

Query Dependent Features: These features aim to capture the relevance of a summary with respect to a query. The features are introduced based on the observation that each summary in the DUC document set is constructed based on a query. We use the frequency of query terms in the summary (after stop word removal) as a feature and the cosine similarity between the summary and the query as another feature.

$$QuerySim(S) = \cos(Query, Summary)$$

NE Frequency: Based on the observation that named entities reflect salient information of the content, we define a *NEScore* as follows as a feature to weight named entities in the summary by their importance in the corpus.

$$NEScore\ \ (S) = \sum_{\forall w} Count\ \ (w_d) \times Count\ \ (w_s)$$

Where $Count(w_s)$ is the frequency of the named entity w in the summary and $Count(w_d)$ is the frequency of the named entity w in the document corpus.

N-Gram Features: Banko and Vanderwende (2004) demonstrated the importance of N-gram in summarization. Their experiments showed that when writing multi-document summaries, human summarizers do not appear to be cutting and pasting phrases in the extractive fashion. On average, they are borrowing text around the bigram level. To mimic the similar behaviour, we extract all the unigrams and bi-grams from the summary and document set. The n-gram similarity between the summary and the document set is used as a feature.

Coherence Features

The coherence features are designed to quantitatively measure the degree to which a sequence of sentences represents a logical flow of thoughts. The Latent Semantic Analysis (Landauer and Dumais, 1997) based features are a coarse statistical measure of this, and measure the drift in content/meaning from one sentence to another.

The semantic space is based on the DUC document collection for each query. The first 15 dimensions are used. LSA is based on singular value decomposition, a mathematical matrix decomposition technique that represents the contextual-usage meaning of words by applying statistical analysis on a large corpus of text. In the reduced semantic space, each sentence in a summary is represented as a vector in this space; similarity between sentences can be computed using cosine distance. We use a sentence to sentence comparison technique whereby a summary of n sentences results in $n-1$ cosine comparisons between the sentences. The *mean* and *standard deviation* of these cosine similarity scores are used as the features.

$$Sim = cos\ (sentence_j, sentence_{j+1})$$

Topic Features

Topic features serve as a basis for evaluating topical relevance of a summary to the document set. The goal is to find the overlap of topics included in a summary and the topics embodied in the document set. Latent Dirichlet Allocation (LDA) technique (Blei et al., 2003; Arora et al., 2008) is used to achieve this goal.

LDA is a generative model for documents, which can be viewed as representing each document as a mixture of topics (represented by a probability distribution over topics). These topics in turn are further represented as a mixture of words. Thus in the context of text modeling, the topic distribution provides an underlying representation of the documents and can be useful in evaluating the summaries.

Gibbs sampling (Griffiths, 2002) is used for inference in the topic model with concentration parameters $\Lambda = 0.1$ and $\beta = 0.01$. We generate 20 topics for each document set and top 100 words from each topic are considered. These top 100 words for each topic are used to calculate the topic

similarity. The detailed algorithm is composed of the following several steps:

1. Run the LDA Model with the number of topics $K=20$ and Gibbs sampling parameter values $\Lambda=0.1$ and $\beta=0.01$.
2. Obtain the *Topic-Word Matrix* from step 1 representing the topic distribution of each word in the vocabulary, i.e., $P(T_i| w_k)$.
3. Generate the *Topic-Summary Distribution* assuming all the words in a summary are independently and identically distributed. This matrix can be obtained by computing the topic probability for each summary, i.e., $P(T_i| S)$, as follows:

$$P(T_i \mid S) = \prod_{k=1}^{N} p(T_i \mid w_k)$$

We then use the computed *Topic-Summary Distribution* for all topics as a feature.

The Evaluation Model

The evaluation model is trained based on the pair wise classification of the summaries where higher pair wise classification accuracy ensures a better summary ranking. We first transform the above-mentioned features into a standard vector notation. Each summary S_i is represented by a feature vector $F= \{f_1, f_2, ..., f_n\}$ where n is the number of features extracted for a particular summary. The training consists of pair wise summaries represented as (F_i, F_j), where F_i and F_j are features of summary S_i and S_j respectively. Then the model is trained using Support Vector Machines in order to learn the relative weights of the features described above. For example, if we use the Responsiveness score for comparison, then the label is 1 if $R_1 > R_2$ where two summaries S_1 and S_2 have responsiveness scores R_1 and R_2 respectively, otherwise 0 if $R_1 < R_2$. We use the LIBSVM package[1] for training and testing. All the parameters are set to their respective default values. That is, SVM type is set to C-SVC and regression function is used as kernel function.

Experiments

Data Preparation

For the model development we used the DUC multi-document summarization dataset. It consists of 50 document sets or queries; each query is accompanied by a collection of relevant documents (25 to 50 documents) from which the summary is to be generated. There are 4 to 9 human written and around 30-40 machine generated summaries in each document set.

We extract the above-mentioned features from these human written and machine generated summaries. DUC also provides evaluations based on

- Human judged linguistic quality, and
- Human judged responsiveness to the query

Our automated evaluation uses both of these scores. For the training purpose 35 of the 50 document sets were used. The remaining 15 document sets were used for testing.

Metrics

Responsiveness score is assigned by the NIST assessors to each human and machine generated summary. This score is a coarse ranking of the summaries for each query, according to the amount of information in the summary that helps to satisfy the information need expressed in the query. The score was an integer between 1 and 5, with 1 being least responsive and 5 being most responsive. In figure 1, we can observe that of the approximately 250 summaries with responsiveness score of 5, around 220 are human written summaries.

Linguistic Quality score is assigned by NIST assessors to each summary for linguistic quality. Five quality questions were used. The linguistic qualities measured were

- Q1: Grammaticality
- Q2: Non-redundancy
- Q3: Referential clarity
- Q4: Focus
- Q5: Structure and coherence

Figure 1 shows the histogram of different metrics for each human score of 1 to 5 for both human written and machine generated summaries.

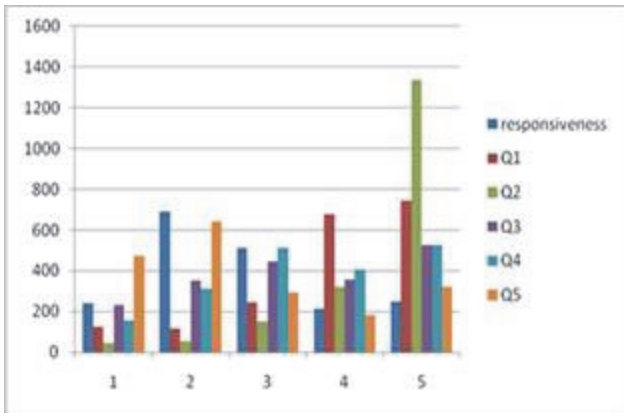

Figure 1: Histogram of different metrics for all summaries (The x-axis represents the human score of 1 to 5 and the y-axis shows the number of summaries assigned to a particular score under different metrics).

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Experimental Results

The experiments are divided into two categories for each metric, namely "with ties" and "without ties". Two summaries, S_1 and S_2, are said to be tied if the human scores for S_1 and S_2 are equal. For our set of summaries under the "without ties" category, we removed the pairs for which the human scores are tied. The different categories shown in Table 1, for example, "Human" refers to the set of summaries only written by the humans; "Human and machine" refers to the set which contains both human written and machine generated summaries. In the "Human vs. Machine" category we try to evaluate how well our system can differentiate between a human written summary and a machine-generated summary.

Baseline

The LSA coherence features were used in the baseline model. These features were extracted from both the human and machine generated summaries. Table 1 shows that the usage of just LSA based features does not perform well in the classification of the summaries.

Table 1: Evaluation Results on Responsiveness Score (shown as a percentage)

Category	With Ties	Without Ties
Human	54.6	56.8
Human and Machine	56.7	59.1
Human vs. Machine	79.1	82.4
Machine	52.3	53.1
Baseline (LSA Coherence Features only)	35.4	37.3
Informativeness Features	55.3	56.1
LSA Coherence Features + Topic features	53.6	54.2
The SVM Regression Model	61.2	64.3
Human and Machine (trained using feature vector triples $(F_i, F_j, F_i\text{-}F_j)$	59.7	63.2

Feature Validation

We have divided the features broadly into three subcategories, Coherence, Topical Relevance and Informativeness. The Coherence category includes the LSA based features; The Topic category includes the *Topic-Summary distribution* feature, and the Informativeness category includes features like n-gram similarity, pseudo summary similarity, Sumbasic, N-gram similarity, etc. Rows 5, 6 and 7 of Table 1 show the accuracy for using these three feature categories independently. Both human written and machine generated summaries ("Human and Machine") were used for this validation. From the experimental results, we find that the informativeness based features work better than the coherence based features and topic features. But the combination of three feature categories outperforms others (59.1 on Human and Machine without ties vs. 56.1 for Informative Features, 37.3 for Coherence features, and 54.2 for Coherence features plus Topic features). The classification accuracy is further improved to 59.7 and 63.2, with and without ties respectively, when trained on feature vector triples $(F_i, F_j, F_i\text{-}F_j)$ in human written and machine generated summaries by combing features in three categories, where $F_i\text{-}F_j$ represents the vector difference between F_i and F_j. The consistent result of 57.3 and 59.4 on Human and Machine, with and without ties respectively, has also been observed by using a training set comprised of a small fraction of the data set (trained on 15 document sets and tested on 35 documents set).

A Support Vector Regression Comparison Model

We also formulate this task as a regression problem where we use a support vector regression model to predict the scores for each summary, which are then used to classify the summary pairs. As shown in Table 1, the regression method outperforms the alternative classification methods and shows that the proposed features are able to learn the goodness of the summary close to human score.

Results Analysis

Table 1 and Table 2 illustrate the experimental results based on human responsiveness scores and linguistic question scores respectively. As observed from Table 1 when comparing row 1 and row 2, the accuracy of pair wise classification of only human written summaries is outperformed by the classification of machine generated summaries and human written summaries, thereby illustrating the relatively lower quality of machine generated summaries. This result is also suggested by the fact that the classification model differentiates well between the human written summaries and the machine generated summaries. The model has an accuracy of as high as 82.4% for summaries without ties to distinguish between the high quality human summary and low quality machine summary. Roughly 30% of the summaries were left after removing ties.

Table 2 shows classification accuracy when trained using different linguistic quality scores. For example, for Linguistic Quality Q2, non-redundancy, the "without ties" accuracy is significantly more than that with ties. From Figure 1, we can observe that most of the summaries have a human score of 5. Thus for the summaries without ties, the model is able to learn the difference between redundant and non-redundant summaries for this question.

The classification using the Linguistic Quality achieves best when we use the Linguistic Quality Q5, Structure and Coherence, with accuracy of 61.4 and 70.8, with and without ties, respectively. The Q5 works best because of the coherence features included in the model. Linguistic Qual-

ity Q3, referential clarity, is second with an accuracy of 60.1 and 63.6 with and without ties, respectively.

Table 2: Evaluation Results on Linguistic Quality

Linguistic Quality	With Ties	Without Ties
Q2	51.7	64.8
Q3	60.1	63.6
Q4	56.2	56.7
Q5	61.4	70.8

Conclusions

In this paper we have presented a new model for stand-alone evaluation of multi-document summaries. This evaluation method can be used within automatic summary generation systems where there are multiple candidate summaries to be ranked. Central to this paper is the extraction of various features covering coherence and informativeness of each summary, and subsequently the usage of these features to rank the summaries without comparison to any reference or human written summary. We also incorporate features such as topic-summary distribution to see how well the summary captures the underlying topics of a document set. Experiments show that this is a promising direction for automatic evaluation of the summaries without the use of a reference summary.

As future work, we believe there is still room for improving the query similarity component in the model through incorporating more query-dependent features. Extending the use of topic models to include the queries themselves may also result in better accuracy in the evaluation model.

Acknowledgements

This research work is supported in part by the NSF grant (IIS-1452898).

References

Papineni, K., SalimRoukos, W., and Zhu, W. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation, *In Proceedings of 2012 Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311-318.

Lin, C. 2004. Rouge: A package for automatic evaluation of summaries. *In Proceedings of the ACL-04 Workshop on Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain.

Lin, C., and Hovy, E. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. *In Proceedings of 2003* Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies *(HLT/NAACL)*, pp. 71–78.

Barzilay, R., and Lapata, M. 2005. Modeling local coherence: An entity-based approach. *In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 141-148.*

Rahimi, Z., Litman, D., Wang, E., and Correnti, R. 2015. Incorporating Coherence of Topics as a Criterion in Automatic Response-to-Text Assessment of the Organization of Writing, *in Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications, 2015*, pp. 20-30.

Higgins, D., and Burstein, J. 2006. Sentence similarity measures for essay coherence. *In Proceedings of the seventh international workshop on computational semantics (IWCS-7)*, Tilburg, The Netherlands, pp. 77-88.

Graesser, C., McNamara, S., and Louwerse, M. 2003. What do readers need to learn in order to process coherence relations in narrative and expository text. *In Sweet, A. P. and Snow, C. E. (Eds.), Rethinking reading comprehension* (pp. 82-98). New York: Guilford Publications.

Haghighi, A., and Vanderwende, L. 2009. Exploring Content Models for Multi-Document Summarization, *in Proceedings of 2009 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, Boulder, Colorado, pp. 362-370.

Jagarlamudi, J., Pingali, P., and Varma, V. 2006. Query Independent Sentence Scoring approach to DUC 2006, in *Document Understanding Conference, 2006 at Annual meeting of HLT/EMNLP*, report No: IIIT/TR/2008/78.

Lau, H, Newman, D., and Baldwin, T. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, 2014.

Chen, H., Branavan, K., Barzilay R., and Karger, R. 2009. Content Modeling Using Latent Permutations, *Journal of Artificial Intelligence Research*, Volume 36, page 129-163, 2009.

Yang, G., Wen, D., Kinshukb, Chen, N., and Sutinen, E. 2015. A novel contextual topic model for multi-document summarization. *Expert Systems with Applications*, vol. 42, pp. 1340-1352.

Nenkova, A. and Venderwende, L. 2005. The Impact of frequency on summarization. *Technical Report*, Microsoft.

Banko, M., and Vanderwende, L. 2004. Using n-grams to understand the nature of summaries. *In Proceedings of the 2014 North American Association for Computational Linguistics (HLT-NAACL)*, pp. 1-4.

Landauer, K., and Dumais, T. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211-240.

Arora, R., and Ravindran, B. 2008. Latent dirichlet allocation based multi-document summarization. *In Proceedings of the SIGIR Workshop on Analytics for Noisy Unstructured Text Data*, pp. 91-97, 2008.

Griffiths, T. Gibbs sampling in the generative model of Latent Dirichlet Allocation. Technical report, Stanford University, 2004.

Foltz, W., Kintsch, W., and Landauer, K. 1998. Textual coherence using latent semantic analysis. *Discourse Processes*, 25(2&3):285–307.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.