Experiments in DUC 2005

Maheedhar Kolla
School of Computer Science
University of Waterloo, Canada
mkolla@cs.uwaterloo.ca

Yllias Chali
Department of Computer Science
University of Lethbridge, Canada
chali@cs.uleth.ca

Abstract

Document Understanding Conference (DUC), organized by NIST, is an evaluation series for automatic summarization systems. In this paper, we give a brief overview of our summarization system, which took part in DUC 2005 evaluation workshop. The objective of this year's workshop is to model a real-world information need, and parallely focus on development of a stable and reliable evaluation procedure.

1 Introduction

Document Understanding Conference(s), DUC, organized by NIST [1], provide a framework to evaluate the system-generated summaries. NIST provides the participants with the document collection(s) and also establishes the guidelines to carry out the summarization task. This year's task models the real-world application in which the user would be interested in learning about a sequence of events. Since the level of interest is a factor in summarization (Sparck-Jones, 1999), this year's task has the user's interest explicitly given to generate a user's profile.

The description of the task is: *given a topic, a user's profile and a collection of documents judged relevant, create a fluent summary (<= 250 words) responding to the information request in a manner specific to the user's profile.* Each participant was provided with 50 topics, with each topic having a minimum of 35 relevant documents. Our participation in DUC 2005 is to experiment with the usage of Information Retrieval techniques to consider only a subset of the document collection for summary generation. Also, we were interested in the evaluation task using Pyramid method.

[1] National Institute of Standards and Technology

This paper is organized as follows: in the following section, we give a brief overview of our system. We then discuss about the various measures and approaches used in the evaluation task.

2 System Overview

The system used in DUC 2005 is an extension of the system used in DUC 2004 (Chali and Kolla, 2004). We give brief overview of each module in the system.

2.1 Pre-processing

In this module, we extract the text from the source document collection and tokenize the text using OAK system[2]. Tokenized text is then segmented into smaller portions using C99 segmentator (Choi, 2000). We then index the segments using Lucene Indexer[3].

2.2 Query Processing

Given a query of the form:

> <Topic> <title> Development of Magnetic Levitation (MAGLEV) Rail Systems </title>
>
> <narr> In what countries are MAGLEV rail systems being proposed?
> Are the proposals for short or long haul?
> Is government financing required for construction?
> </narr>
>
> <granularity> specific </granularity>
>
> </Topic>

We identify the noun phrases (NPs) from the title and the narrative portion of the topic. For the above example, the query terms extracted are:

[2] http://nlp.cs.nyu.edu/oak/
[3] http://lucene.apache.org/

maglev rail systems short or long haul countries
government financing proposals
magnetic levitation (maglev) rail systems construction development

Now, we query the indexed document collection, with the above extracted query terms and retain the top 50 segments. This would allow us to filter the portions of the document, which are not relevant to the topic. On the other hand, there is a possibility that we lose some segments which are relevant but which do not have an occurrence of the query term.

2.3 Clustering

Extracted segments are grouped into clusters, based on their topical similarity as explained in (Kolla, 2005). We first compute the lexical chains for each segment and then compute the similarity between the segments based on the lexical chain overlap. Segments are then retained in the cluster, in which they contribute the most.

2.4 Extraction

Once the clusters are generated, we rank the clusters based on the *tf.idf*() value (Salton and Buckley, 1987) of the query terms. Given query terms Q(1..n), score of a cluster C_j can be computed as:

$$score(C_j) = \sum_{i=1}^{n} tf(term_{ij}), idf(term_i) \quad (1)$$

where
$tf(term_{ij})$ - is the frequency of the query term Q_i in C_j.
$idf(term_i)$ - is the *idf* value of the term Q_i.

We then calculate the score of segments and sentences within the cluster(s) in the similar way. Summary is then generated by extracting the top-ranked *n* sentence(s) from each cluster, i.e., first ranked sentences from all clusters, followed by second ranked ones and so on, until the length of the summary required is reached.

3 Evaluation

One of the goal(s) of this year's DUC workshop is to develop evaluation measures which consider the meaning of the words in the given context towards its importance. Three different schemes have been used to evaluate the system-generated summaries:

- Human evaluation.
- ROUGE automatic evaluation.
- Pyramid method.

System	ScaleResp	allResp	rawResp	Avg.Qual
26	12.35	12.52	2.06	3.18
Baseline	12.61	12.68	1.98	4.41
Avg.Hum	N/A	35.26	4.67	4.86
Avg.Sys	16.63	16.82	2.40	3.26

Table 1: Responsiveness and Linguistic Quality Measures

3.1 Human Evaluation

NIST judges carried out the manual evaluation of the summaries to measure the responsiveness (relative) and linguistic quality. Responsiveness can be defined as the measure of the extent to which the summary is able to satisfy the information need of the user, relative to the others. Each summary is assigned a value between 1 and 5, where 5 being the best. NIST also evaluated the linguistic quality of the summaries. Each judge was asked to determine the readability, grammatical correctness etc. of the summaries, evaluated independent of the model summaries.

- Grammaticality
- Non-redundancy
- Referential clarity
- Focus
- Structure and coherence.

Each summary is judged for each of the above linguistic quality and is given a value from 1 to 5, where 1 is the best. Table 1 shows the responsiveness and linguistic quality of our system.

3.2 ROUGE

ROUGE, Recall-Oriented Understudy of Gisting Evaluation, is an automatic method of evaluation based on the *n-gram*, where n=1,2,3,4 overlaps between the system-generated and the model summaries (Lin, 2004). ROUGE measures have been used in previous DUC (Over, 2004) evaluation. (Lin, 2004) found that ROUGE method evaluation correlates strongly with that of the human evaluation. In this year's evaluation, ROUGE-2 and ROUGE-SU4 have been used as the official measures. Table 2 shows the results of the ROUGE evaluation.

Even though ROUGE provides an automatic method to evaluate the systems, in comparison with the human summaries, a study (Nenkova and Passonneau, 2004) showed that ROUGE measures cannot be used as an absolute measure of the system's performance. They proposed a

System	ROUGE-SU4	ROUGE-2
26	0.10	0.05
Baseline	0.09	0.04
Avg.Hum	0.16	0.10
Avg.Sys	0.11	0.06

Table 2: ROUGE Evaluation

method to evaluate summaries based on the content overlap among a pool of human summaries rather than one model summary.

3.3 Pyramid Evaluation

(Nenkova and Passonneau, 2004) defined pyramid as a weighted inventory of Summarization Content Units, SCU's. A SCU can be defined as the smallest unit of a sentence, almost a clause length, and which refer to some semantic meaning. Each SCU has a *label* and *a set of contributors*. The contributors are pieces of text, identified from the pool of human summaries, each referring to the same meaning in the given context. For example:

> *Clinical trials are performed when a new drug is developed.*
>
> A1. The second phase is clinical trials
> B1. a potential drug goes into clinical trials
> C1. drugs are tested for safety on healthy human volunteers
> D1. then with clinical trials on humans.

For the SCU above, there are four contributors across the pool of human summaries. Once all the sentences in the pool of human summaries are annotated, a pyramid is constructed comprising *n*-Tiers of SCU's. SCU's belonging to one tier have the same weight, which is equal to the number of contributors for that SCU, in the human summaries. Given a pyramid, an ideal summary(ies) would then be considered as those which have most of the SCU's belonging to the top tier(s).

In DUC 2005 , NIST judges created 9 summaries for a subset of 20 query topics. 27 participants volunteered to carry out the evaluation exercise, which was co-ordinated by Columbia University[4]. Each participating group was assigned with one topic, six of the topics (324, 400, 407, 426, 633 and 695) were assigned to more than one group. To maintain consistency, all summaries for a particular topic are supposed to be annotated by one designated person. Also, they provided the groups with some consistency checking scripts to maintain a consistency in annotating the same sentence, occurring in different system-generated summaries.

On completion of the annotation process by the participant(s), any errors in the annotation procedure were corrected by the Columbia University group people. Two different score(s) could be computed in this evaluation process. One is *score* of a peer summary with *m* SCU's. It is equal to the sum of the weights of the SCU's present in the peer summary divided by the weight of an ideal summary. An ideal summary, with *m* SCU's, can be defined as the summary which has all the SCU's from the top *n*th tier before having any SCU's from the *n-1*th tier.

The second measure, *modified score*, is the ratio of the ideal weight of the peer summary, with content size equal to the content size of an average human summary, used in the construction of the pyramid. From the discussion[5], it can be understood that the first measure can be used as the precision and the second corresponds to the recall value of the summaries. Table 3 shows the average precision and average recall measures computed for all of the systems, which have taken part in pyramid evaluation.

4 Discussion and Conclusion

In this paper, we gave a brief description of our summarization system, in context of DUC 2005 participation. We also briefed the various evaluation measures carried out in this year's workshop. We were unable to experiment with the BE package[6], and would like to carry on with experiments using that package to see the co-relation between that approach and the others. Also, it would be interesting to see if we could project the system's performance based on the 20 test topics used in the pyramid evaluation.

In this year's participation, we assumed that the query terms are independent of each other and hence converted them to their baseform before querying. This approach would not work for complex queries, which require more deeper analysis to obtain the user's need.

5 Acknowledgments

We would like to thank Ali Karim, a co-op student for his work in the development of the system. Also, we would like to thank the NLP group at Columbia University for organizing the pyramid evaluation and also for providing both processed and un-processed peer annotation files. We would also like to thank Guy Lapalme, for the analysis of the ROUGE and responsiveness evaluation results.

[4] www.cs.columbia.edu/nlp

[5] in DUC mailing list

[6] www.isi.edu/ cyl/BE/

System	AvgP	AvgR
3	0.189	0.145
4	0.218	0.171
6	0.212	0.165
7	0.230	0.184
10	0.232	0.198
11	0.208	0.168
12	0.207	0.164
13	0.186	0.141
14	0.246	0.186
15	0.229	0.177
16	0.217	0.171
17	0.238	0.196
19	0.208	0.167
20	0.146	0.096
21	0.207	0.160
23	0.123	0.061
24	0.147	0.113
25	0.168	0.138
26	0.197	0.140
27	0.164	0.132
28	0.194	0.139
30	0.136	0.112
31	0.156	0.119
32	0.213	0.160
A	0.488	0.461
B	0.542	0.481
C	0.705	0.666

Table 3: Avg Precision and Avg Recall for all Systems

References

Y. Chali and M. Kolla. (2004). Summarization techniques at DUC 2004. In *Proceedings of the Document Understanding Conference*, pages 105 -111, Boston. NIST.

F. Y. Y. Choi. (2000). Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics*, pages 26 - 33, Seattle, Washington.

M. Kolla. (2005). Automatic text summarization using lexical chains: Algorithms and experiments. Master's thesis, Department of Computer Science, University of Lethbridge.

C. Y. Lin. (2004). Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74 - 81, Barcelona, Spain.

A. Nenkova and R. Passonneau. (2004). Evaluating content selection in summarization: the Pyramid method. In *Proceedings of the Human Language Technology Research Conference/North American Chapter of the Association of COmputation Linguistics*, pages 145-152, Boston, MA.

P. Over. (2004). Introduction to DUC 2004: an intrinsic evaluation of generic news text summarization systems. In *Proceedings of the Document Understanding Conference*, Boston MA.

G. Salton and C. Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, USA.

K. Sparck-Jones. (1999). Automatic summarizing: Factors and directions. In Mani and Maybury, editors, *Advances in Automatic Text Summarization*. MIT press.