Efficient Randomized Algorithms for Text Summarization

Ahmed Mohamed and Sanguthevar Rajasekaran

Department of Computer Science & Engineering,
University of Connecticut, Storrs, CT 06268
{amohamed, rajasek}@engr.uconn.edu

Abstract. Text summarization is an important problem since it has numerous applications. This problem has been extensively studied and many approaches have been pro-posed in the literature for its solution. One such interesting approach is that of posing summarization as an optimization problem and using genetic algorithms to solve this optimization problem. In this paper we present elegant randomized algorithms for summarization based on sampling. Our experimental results show that our algorithms yield better accuracy than genetic algorithms while significantly saving on time. We have employed data from Document Understanding Conference 2002 and 2004 (DUC-2002, DUC-2004) in our experiments.

1 Introduction

Document summarization has been the focus of many researchers for the last decade, due to the increase in on-line information and the need to find the most important information in a (set of) document(s). There are different approaches to generate summaries depending on the task the summarization is required for. Summarization approaches usually fall into 3 categories (Mani and Maybury, 1999):

- *Surface-level* approaches tend to represent information in terms of shallow features, which are then selectively combined together to yield a salience function used to extract information;
- *Entity-level* approaches build an internal representation for text, modeling text entities and their relationships. These approaches tend to represent patterns of connectivity in the text (e.g., graph topology to help determine what is salient);
- *Discourse-level* approaches model the global structure of the text, and its relation to communicative goals.

Some approaches mix between two or more of the features of the above mentioned approaches, and the approaches discussed in this paper fall in that category, since they involve both surface and entity levels' features.

© *A. Gelbukh (Editor)*
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 195-200

Acknowledgments This project was funded by Bell University Laboratories (BUL) and the Canada Natural Science and Engineering Research Council (NSERC). The authors would also like to thank to anonymous referees for their comments.

References

1. Green, B.F., Wolf, A.K., Chomsky, C., Laughery, K.: Baseball: An Automatic Question Answerer. In: Proceedings of the Western Joint Computer Conference. (1961) 219–224
2. Woods, W.: Progress in Natural Language Understanding: An Application to Lunar Geology. In: AFIPS Conference Proceeding. Volume 42. (1973) 441–450
3. Mollá, D., Vicedo, J.L., eds.: Workshop on Question Answering in Restricted Domains - ACL-2004, Barcelona, Spain (2004)
4. Mollá, D., Vicedo, J.L., eds.: AAAI-05 Workshop on Question Answering in Restricted Domains, Pittsburg, Pennsylvania (2005) to appear.
5. Mollá, D., Vicedo, J.L., eds.: Special Issue of Computational Linguistics on Question Answering in Restricted Domains. (2005) to appear.
6. Herzog, O., Rollinger, C.R., eds.: Text Understanding in LILOG, Integrating Computational Linguistics and Artificial Intelligence, Final Report on the IBM Germany LILOG-Project. In Herzog, O., Rollinger, C.R., eds.: Text Understanding in LILOG. Volume 546 of Lecture Notes in Computer Science., Springer (1991)
7. Mollá, D., Berri, J., Hess, M.: A real world implementation of answer extraction. In: Proceedings of the 9th International Workshop on Database and Expert Systems, Workshop: Natural Language and Information Systems (NLIS-98), Vienna (1998)
8. Benamara, F., Saint-Dizier, P.: Advanced Relaxation for Cooperative Question Answering. In: New Directions in Question Answering. (2004) 263–274
9. Oroumchian, F., Darrudi, E., Ofoghi, B.: Knowledge-Based Question Answering with Human Plausible Reasoning. In: Proceedings of the 5th International Conference on Recent Advances in Soft Computing. (2004)
10. Light, M., Mann, G., Riloff, E., Breck, E.: Analyses for Elucidating Current Question Answering Technology. Natural Language Engineering **7** (2001)
11. Diekema, A.R., Yilmazel, O., Liddy, E.D.: Evaluation of Restricted Domain Question-Answering Systems. In: Proceedings of the Association of Computational Linguistics 2004 Workshop on Question Answering in Restricted Domains (ACL-2004), Barcelona, Spain (2004)
12. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: OKAPI at TREC-3. In D.K.Hartman, ed.: Overview of the Third TExt Retrieval Conference (TREC-3), Gaithersburg, NIST (1995) 109–130
13. Doan-Nguyen, H., Kosseim, L.: Improving the Precision of a Closed-Domain Question-Answering System with Semantic Information, Avignon, France, RIAO-2004 (2004)
14. Doan-Nguyen, H., Kosseim, L.: The Problem of Precision in Restricted-Domain Question-Answering. Some Proposed Methods of Improvement, Barcelona, Spain, ACL (2004)
15. Kowalski, G.: Information Retrieval Systems – Theory and Implementation. Kluwer Academic Publishers, Boston/Dordrecht/London (1997)

Efficient Randomized Algorithms for Text Summarization

Ahmed Mohamed and Sanguthevar Rajasekaran

Department of Computer Science & Engineering,
University of Connecticut, Storrs, CT 06268
{amohamed, rajasek}@engr.uconn.edu

Abstract. Text summarization is an important problem since it has numerous applications. This problem has been extensively studied and many approaches have been pro-posed in the literature for its solution. One such interesting approach is that of posing summarization as an optimization problem and using genetic algorithms to solve this optimization problem. In this paper we present elegant randomized algorithms for summarization based on sampling. Our experimental results show that our algorithms yield better accuracy than genetic algorithms while significantly saving on time. We have employed data from Document Understanding Conference 2002 and 2004 (DUC-2002, DUC-2004) in our experiments.

1 Introduction

Document summarization has been the focus of many researchers for the last decade, due to the increase in on-line information and the need to find the most important information in a (set of) document(s). There are different approaches to generate summaries depending on the task the summarization is required for. Summarization approaches usually fall into 3 categories (Mani and Maybury, 1999):

- *Surface-level* approaches tend to represent information in terms of shallow features, which are then selectively combined together to yield a salience function used to extract information;
- *Entity-level* approaches build an internal representation for text, modeling text entities and their relationships. These approaches tend to represent patterns of connectivity in the text (e.g., graph topology to help determine what is salient);
- *Discourse-level* approaches model the global structure of the text, and its relation to communicative goals.

Some approaches mix between two or more of the features of the above mentioned approaches, and the approaches discussed in this paper fall in that category, since they involve both surface and entity levels' features.

© *A. Gelbukh (Editor)*
Advances in Natural Language Processing
Research in Computing Science 18, 2006, pp. 195-200

2 Background and Related Work

2.1 ExtraNews

In our study, we considered the work done at LARIS laboratory (Fatma et al., 2004). In this approach (called ExtraNews) summarization is considered as an optimization problem. A set of summaries is generated randomly and then a Genetic Algorithm is utilized to come up with a good summary. They use a fitness function that depends on three different factors. The first factor $\omega 1$ is related to the length of the summary, in which the length of the summary is tested against the required target length, as shown in the following equation:

$$\omega_1 = \begin{cases} \dfrac{\sum_{i=1}^{m} L(ph_i)}{L_E} & \text{if } \sum_{i=1}^{m} L(ph_i) < 0.9 \times L_E \\ 0 & \text{if } \sum_{i=1}^{m} L(ph_i) > L_E \end{cases} \tag{1}$$

where $L(ph_i)$ is the length of sentence i; L_E is the target summary length set by the user and m is the number of sentences in the summary.

The second factor ω_2 pertains to the coverage criterion. It calculates how many of the original keywords have been captured in the target summary:

$$\omega_2 = \frac{\sum M_{ext}}{\sum M_{doc}} \tag{2}$$

where M_{ext} represents the keywords in the summary and M_{doc} represents the keywords in the source document-set.

The last factor ω_3 is associated with the weight criterion. It is the fraction of the sum of weights of all sentences in the summary to the maximal summary weight in the population:

$$\omega_3 = \frac{\sum P_{ext}}{Max(P_{pop})} \tag{3}$$

where P_{ext} is the weight of a sentence of the summary and $Max(P_{pop})$ is the maximum summary weight in the population. It was not mentioned, however, in (Fatma et al., 2004) how the weight of the sentence is calculated. For this reason, we choose to use a cosine similarity measure (Salton et al., 1997) to weight each sentence, in which the summary and each sentence in the summary are represented as vectors of terms and then the weight is calculated from the following formula:

$$sim(D1, D2) = d1 \bullet d2 \tag{4}$$

where: D_1, D_2 are documents 1 and 2 respectively, and d_1, d_2 are the term vectors of documents 1 and 2 respectively..

It is also important to note that (Fatma et al., 2004) did not mention how the above three coefficients were composed together to form the fitness function. So, we assumed that the fitness function is the product of all coefficients.

ExtraNews system ranked very well in tasks 4 (creating a short summary for English translations of a document cluster) and 5 (creating a short summary from a document cluster answering the question "Who is X?" where X is a name of a person/Group of people) of the Data Understanding Conference (DUC-2004) tasks. It ranked above average in task 2 (creating a short summary for each document cluster) and in task 3a (creating a very short summary for automatic English translation of a document cluster). However, it ranked badly in tasks 1 (creating a very short summary for English translation of a document cluster) and 3b (creating a very short summary for manual English translation of a document cluster) (Over, 2002). They attributed the last bad results to the fact that the phrases considered in the segmentation process are not very well suitable for very short summaries.

2.2 Our Randomized Algorithms

Randomized algorithms have played a vital role in the past three decades in solving many fundamental problems of computing efficiently. Many problems have been shown to be better solvable using randomization than determinism. For examples see (Horowitz, Sahni and Rajasekaran 1998). Algorithms such as simulated annealing that have proven very effective in solving some intractable problems in practice are examples of randomized algorithms. Both in practice and theory randomization has resulted in the design of efficient algorithms.

One popular theme in randomization has been that of sampling. In its simplest form sampling can be defined as follows. Say we want to measure a certain characteristic C from a dataset D. We could do this by processing all the points in D. Alternatively we could pick a random subset D' (called the sample) of D, measure the same characteristic in D', and from this sample measurement infer the value of the characteristic in D. Preferably, we should be able to infer this value with high probability. For a survey of sampling techniques see (Rajasekaran and Krizanc 2001). Our algorithm for summarization is based on sampling. Before presenting our algorithm, we briefly describe the approach taken in the ExtraNews system. This sys-tem employs genetic algorithms.

The genetic algorithm for solving any optimization problem has been motivated by Darwin's theory of evolution and works as follows. A population of random points from the feasible space is chosen at the beginning. The 'fitness' of each point is computed. A new population is obtained from the old one using two operators, namely, crossover and mutation. A crossover operation refers to taking two points in the population and producing an 'offspring' point similar to the way an offspring chromosome is produced from two parent chromosomes. Crossovers are per-formed typically between pairs with high fitness values (with the hope that the offspring will be fitter). The above process of producing a new population from an old one is repeated until certain conditions are satisfied. For example, the algorithm could terminate after producing a certain number of populations or when the best solution in the population does not change significantly (from one population to the next).

Fatma *et al.* have employed genetic algorithms in the context of summarization as follows. They first produce a population of random points. Each point is nothing but a summary formed by random sentences picked. Fitness of each point is calculated. A new population is then produced by using the GA operators (crossover and mutation) from the older population and the newer population replaces the older one, and so on. Every time a new population is formed, the best summary is kept in a safe place until a better summary is found, then the better summary will replace the poorer summary.

We propose a randomized algorithm based on sampling. We pick a random sample as in (Fatma *et al.*) and choose the best summary in this sample. The process stops after two generations only. We employ the three criteria that Fatma *et al.* have employed for measuring fitness of summaries.

It is important to mention that our approach uses the same fitness function used with the GA as a built-in function. So, the GA is not required to run in conjunction with our approach.

3 Data and Experimental Design

3.1 Data

We used multi-document extracts from DUC-2002 and from DUC-2004 (task 2) in our experiment. In the corpus of DUC-2002, each of the ten information analysts from the National Institute of Standards and Technology (NIST) chose one set of newswire/paper articles in the following topics (Over 2002):

- A single natural disaster event with documents created within at most a 7-day window;
- A single event of any type with documents created within at most a 7-day window;
- Multiple distinct events of the same type (no time limit);
- Biographical (discuss a single person);

Each assessor chose 2 more sets of articles so that we ended up with a total of 15 document sets of each type. Each set contains about 10 documents. All documents in a set are mainly about a specific "concept."

The corpus of DUC-2004 (task 2) is composed of 50 TDT English news clusters. Each cluster contains about 10 documents chosen by NIST about one single event (Over and Yen, 2004).

3.2 Experimental Design

A total of 59 document-sets from DUC-2002 and 50 document-sets from DUC-2004 have been used in our experiment to investigate the performance of our randomized algorithm. We ran the Genetic Algorithm (GA) on both corpuses. Since we are comparing our algorithm's performance to that of the GA, we found it more meaningful to use the fitness function (Fatma et al., 2004) used in their GA to evaluate the quality of our summaries as well.

4 Results

Tables 1 and 2 show a comparison between the results obtained using DUC-2002 and DUC-2004, respectively, of GA and our RA performance in terms of quality, number of times summaries produced faster, average quality and average time spent by each algorithm, respectively. The experimental results show that the randomized algorithm produced competitive results in much less time than the Genetic Algorithm.

Table 1. Comparison between GA and RA results from the DUC-2002 data.

	GA	RA
# of best summaries	5/59	54/59
# of summaries produced faster	0/59	59/59
Average quality	0.129	0.152
Average time (sec)	23.1	13.6

Table 2. Comparison between GA and RA results from the DUC-2004 data.

	GA	RA
# of best summaries	8/50	42/50
# of summaries produced faster	0/50	50/50
Average quality	0.041	0.051
Average time (sec)	8.4	3.5

5 Conclusions

In this paper we have presented an elegant randomized algorithm for text summarization. This algorithm is based on sampling. Our algorithm has been compared with the Genetic Algorithm of (Fatma et al. 2004). This comparison shows that our randomized algorithm produces summaries that are comparable in quality to those produced by GA while taking much less time. An important open problem is to study if sampling can be used in conjunction with other text summarization approaches to obtain similar speedups. We are also planning on testing our approach against the GA approach on the DUC-2004 collection when it is ready for experimentation.

Acknowledgment. SR has been supported in part by the NSF Grant ITR: 0326155.

References

1. J. K. Fatma, J. Maher, B. H. Lamia, B. H. Abdelmajid, LARIS Laboratory. 2004. Summarization at LARIS Laboratory. Document Understanding Workshop. May 6-7, 2004, Boston Park Plaza Hotel and Towers, Boston, USA
2. E. Horowitz, S. Sahni and S. Rajasekaran, Computer Algorithms, W. H. Freeman Press, 1998.
3. Chin-Yew Lin and Eduard Hovy. 2002. Manual and Automatic Evaluation of Summaries. In Proceedings of the Workshop on Automatic Summarization post conference workshop of ACL-02, Philadelphia, PA, U.S.A., July 11-12 (DUC2002).
4. Inderjeet Mani and Mark T. Maybury. (1999). Advances in Automatic Text Summarization. The MIT Press.
5. Paul Over and James Yen. 2004. An Introduction to DUC 2004 Intrinsic Evaluation of Generic New Text Summarization Systems. Document Understainding Conferences website (http://www-nlpir.nist.gov/projects/duc/)
6. Paul Over and Walter Liggett. 2002. Introduction to DUC-2002: an Intrinsic Evaluation of Generic News Text Summarization Systems. Document Understand-ing Conferences website (http://duc.nist.gov/)
7. Sanguthevar Rajasekaran and Danny Krizanc, Random Sampling: Sorting and Selection, in Handbook of Randomized Computing, Kluwer Academic Press, 2001.
8. Gerard Salton. 1988. Automatic Text Processing. Addison-Wesley Publishing Company.
9. Gerard Salton, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic Text Structuring and Summarization. Information Processing and Management 33(2): 193-207. Reprinted in Advances in Automatic Text Summarization, I. Mani and M.T. Maybury (eds.), 341–35. Cambridge, MA: MIT Press.