A Prototype Crowdsourcing Approach for Document Summarization Service

Hajime Mizuyama, Keishi Yamashita, Kenji Hitomi, Michiko Anse

A Prototype Crowdsourcing Approach for Document Summarization Service

Hajime Mizuyama, Keishi Yamashita, Kenji Hitomi, Michiko Anse

Aoyama Gakuin University,
5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, Kanagawa 252-5258, Japan
mizuyama@ise.aoyama.ac.jp

Abstract. This paper proposes a crowdsourcing approach for informative document summarization service. It first captures the task of summarizing a lengthy document as a bi-objective combinatorial optimization problem. One objective function to be minimized is the time to comprehend the summary, and the other one to be maximized is the amount of information content remaining in it. The solution space of the problem is composed of various combinations of candidate condensed elements covering the whole document as a set. Since it is not easy for a computer algorithm to create condensed elements of different lengths which are natural and easy for a human to comprehend, as well as to evaluate the two objective functions for any possible summary, these sub-tasks are crowdsourced to human contributors. The rest of the approach is handled by a computer algorithm. How the approach functions is tested by a laboratory experiment using a pilot system implemented as a web application.

Keywords: collective intelligence, crowdsourcing, document summarization, human computation, micro tasks

1 Introduction

There are manual and computerized document summarization services today. Since manual summarization is costly and usually takes a long time, many studies have been conducted on computerized summarization [1][2]. Computerized techniques for document summarization can be classified into extractive and abstractive methods. The extractive method identifies important sentences or phrases in the input document, and outputs a summary by simply connecting the identified sentences or phrases. The abstractive method, on the other hand, newly creates condensed sentences so that their combination can deliver whole relevant information in the original document.

Since the extractive method is easier to automate than the abstractive one, most computerized summarization techniques practically applicable today are classified into the former category. However, a summary obtained by the extractive method has a fragmentary nature, that is, it only covers the fragmentary information contained in the chosen sentences or phrases. Therefore, it is appropriate for indicative purpose but for informative purpose. A summary for indicative purpose is used to determine

adfa, p. 1, 2011.

whether the user should read the original document, whereas a summary for informative purpose gives the user sufficient information to proceed without reading the original document.

Thus, for the purpose of informative document summarization service, the abstractive method should be applied. However, computerized techniques in this category are still in early development phase, especially because it is not easy for a computer algorithm alone to create sentences which are natural and easy for a human to comprehend. In order to overcome this difficulty, this paper takes a crowdsourcing approach. Crowdsourcing is a form of outsourcing a task, where the task is divided into many micro pieces and the pieces are outsourced, typically to a lot of anonymous people with a tiny wage over the internet [3]. This approach is also called human computation [4][5][6] and can be used as if a part of a computer algorithm. Its successful applications include parts classification [7], document translation [8][9][10], and document editing [11].

In the remainder of this paper, after introducing a model of document summarization task, a prototype crowdsourcing approach for accomplishing the task is proposed. Then, how the approach functions is tested by a laboratory experiment using a pilot system implemented as a web application. Following its results, discussion and conclusions are provided.

2 Modeling Document Summarization Task

Any document is composed of several units, for example, chapters, sections, paragraphs, or sentences. In this paper, we distinguish two types of units; one is evaluation units and the other is condensation elements. Then, the whole document to be summarized is captured as a set of evaluation units, and each evaluations unit as a sequence of a manageable number of condensation elements. How to define the scope of these units is not unique, but the scope can be assigned to the input document recursively. For example, chapters can be deemed as evaluation units and sections as condensation elements. It is also possible to define paragraphs as evaluation units and sentences as condensation elements. It is assumed that document summarization task starts from a scope having the smallest condensation elements and gradually shifts to a coarser one as the task proceeds. In the following, we will focus on an evaluation element and provide a model for the task of summarizing it. The overall task of summarizing the whole document can be captured as parallel and recursive application of the modeled task.

The task of summarizing an evaluation unit is captured as a bi-objective combinatorial optimization problem. One objective function to be minimized is the time to comprehend the summary, and the other one to be maximized is the amount of information content remaining in it. The solution space of the problem comprises various combinations of condensed elements covering the whole evaluation unit. For example, suppose that the concerned evaluation unit is composed of an ordered set of condensation elements ($= U_0$), and there are several candidate condensed elements ($\in U_1$). Further, each condensed element m corresponds to an ordered set V_m, which is a

sub-sequence of U_0. That is, condensed element m is an efficient expression of the information contained in V_m. It is also defined that for all $m \in U_0$, V_m equals $\{m\}$. Then, every feasible solution of the summarization problem can be captured as an ordered set S_k, whose elements are taken from $U_0 \cup U_1$, satisfying the following conditions:

$$\cup_{m \in S_k} V_m = U_0 \tag{1}$$

$$V_m \cap V_n = \emptyset \ \ (\forall m, n \in S_k) \tag{2}$$

Let us denote the two objective functions, the time to comprehend S_k and the amount of information content remaining in it, by $F_1(S_k)$ and $F_2(S_k)$ respectively. Then, the summarization problem can be formulated as:

Minimize $F_1(S_k)$ and maximize $F_2(S_k)$

Subject to equations (1) and (2)

It is, however, noted that the objective functions $F_1(S_k)$ and $F_2(S_k)$ as well as the set U_1 are not given a priori, and establishing these is also included in the summarization task.

3 Proposed Crowdsourcing Approach

In this section, we propose a prototype crowdsourcing approach for the modeled document summarization task. An outline of the proposed approach is shown in **Fig. 1**. According to the task model provided above, the summarization task includes three sub-tasks, that is, creation, evaluation, and optimization. These sub-tasks are addressed one by one in the following.

3.1 Creation

As shown in **FIg. 1**, the proposed approach starts with dividing the concerned evaluation unit into elements. Then, the obtained condensation elements are numbered and stored into a database, so that each element can be taken out easily by specifying its number. At this point, the set U_1 is empty. Hence, condensed elements corresponding to various sub-sequences of U_0 should be created and thrown into the set U_1. In the proposed approach, this sub-task is treated as a set of micro tasks, each of which corresponds to the task of creating a single condensed element. Then, the micro tasks are crowdsourced, and their outputs are also numbered and added to the same database.

How each of the micro tasks is performed by a human contributor is as follows. When a contributor starts the micro task, she/he is shown the whole evaluation unit as a sequence of condensation elements. Then, she/he is supposed to choose a subsequence of them and to create a more efficient expression representing the information contained in the sub-sequence. If someone else has already created a con-

densed element corresponding to the same sub-sequence, the element is also shown to her/him as a hint.

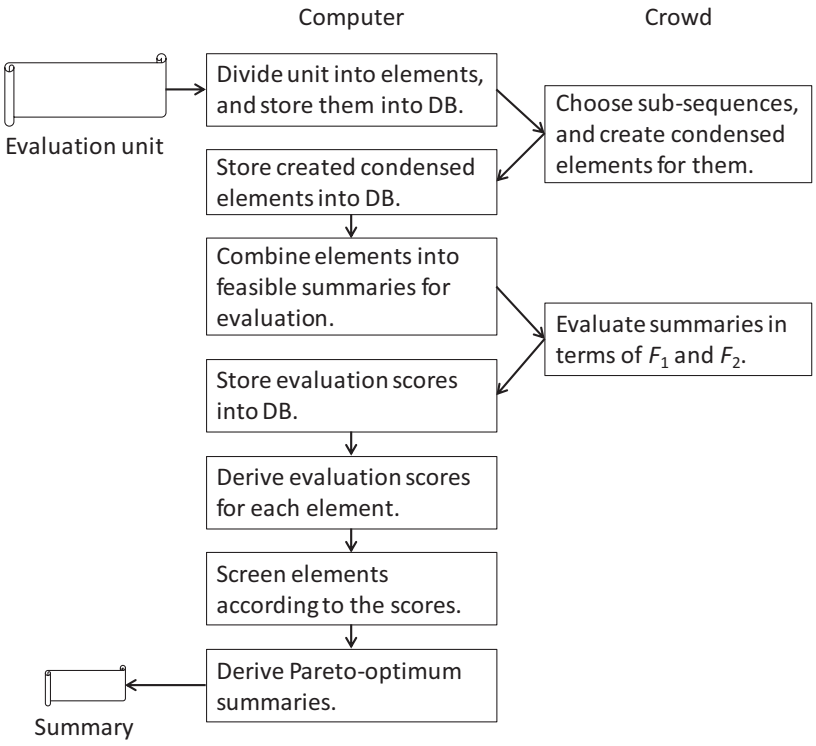

Fig. 1. Outline of proposed approach

3.2 Evaluation

Since the creation sub-task described above enriches the set U_1, the number of feasible combinations of condensed elements, i.e. summaries, will become so large that all of them cannot be evaluated one by one manually. Therefore, numerical expressions for the objective functions $F_1(S_k)$ and $F_2(S_k)$ are necessary. Thus, in the proposed approach, we formulate the evaluation measures as follows:

$$F_1(S_k) = 100 - \sum_{i \in S_k} f_{1i} \tag{3}$$

$$F_2(S_k) = 100 - \sum_{i \in S_k} f_{2i} \tag{4}$$

It is noted that, for simplicity, only main effects are considered in the equations. If the value of $F_1(S_k)$ is 100, the time length required for comprehending the summary is as long as that for comprehending the original evaluation unit. Its value decreases as the time length required for comprehending the summary decreases. On the other hand,

when no relevant information is lost in the summary, the value of $F_2(S_k)$ is 100. Its value decreases as the information remaining in the summary decreases.

In order to estimate the parameter values in equations (3) and (4), simple multiple regression using dummy variables can be utilized. The regression analysis needs some learning data, and hence the data should be gathered somehow. In the proposed approach, this sub-task is treated as a set of micro tasks, each of which corresponds to the task of evaluating a single summary in terms of $F_1(S_k)$ and $F_2(S_k)$. Then, the micro tasks are crowdsourced, and their outputs are stored in a database. Which summaries are to be evaluated is determined by the computer according to a rationale called D-optimality of experimental design.

How each of the micro tasks is performed by a human contributor is as follows. She/he is shown the original evaluation unit and a feasible summary chosen by the computer, and is supposed to read and comprehend the both. She/he is supposed to push a button on a web browser by a computer mouse, when she/he starts and ends reading each of the texts. This makes it possible to quantify the time length required for comprehending each text, and objectively evaluate the value of $F_1(S_k)$ according to the ratio between the quantified time lengths for the summary and the original evaluation unit. She/he is also asked to subjectively evaluate the amount of information contents remaining in the summary with a score from 0 to 100. This score can be used as a sample value of $F_2(S_k)$.

3.3 Optimization

When a sufficient number of condensed elements are supplied by the creation sub-task, a sufficient number of learning data are obtained by the evaluation sub-task, and numerical expressions are derived for the objective functions $F_1(S_k)$ and $F_2(S_k)$ through multiple regression analysis using dummy variables, then we can proceed to the optimization sub-task. At this point, this sub-task can be captured as a simple bi-objective combinatorial optimization problem. When the solution space is large, various meta-heuristics can be utilized.

However, in the following laboratory experiment, a simple two step approach is taken, since the problem size is not so large. At the first step, non-Pareto-optimum condensed elements are screened out for each sub-sequence of the evaluation unit. Then, at the second step, Pareto-optimum summaries are chosen from the whole possible combinations of the remaining condensed elements.

4 Laboratory Experiment

4.1 Implementation

In this section, how the proposed approach functions is tested with a small-scale laboratory experiment. To conduct the experiment, the prototype crowdsourcing approach for informative document summarization is implemented as an elementary web application. The application uses MySQL as the database storing the condensa-

tion and condensed elements as well as the sample evaluation scores. It also uses PHP for handling various interactions with human contributors, and R for deriving the experimental design for the evaluation sub-task and conducting multiple regression analysis using dummy variables.

4.2 Outline of Experiment

The laboratory experiment comprises three phases. The first and second phases test whether the creation sub-task and the evaluation sub-task function properly. The third phase investigates the quality of the output summaries. In the experiment, we use a Japanese document on global warming having three paragraphs and 833 characters as the input evaluation unit, and treat its paragraphs as the condensation elements.

In the first phase, six male senior students of Aoyama Gakuin University participated in the experiment as contributors. The creation sub-task was performed by them using the developed web application, until at least three condensed elements have been obtained for every possible sub-sequence of the evaluation unit, that is, {1}, {2}, {3}, {1, 2}, {2, 3}, and {1, 2, 3}. As a result, it is confirmed that the proposed approach can actually collect various candidate condensed elements from multiple human contributors.

In the second phase, four male senior students of the same university participated. They as a whole have evaluated fifty summaries specified by the computer using the web application. Further, the parameter values of equations (3) and (4) were successfully estimated using the obtained evaluation scores as the learning data for multiple regression analysis. As a result, it is shown that the proposed approach can collect sufficient learning data in practice for establishing objective functions $F_1(S_k)$ and $F_2(S_k)$ from multiple human contributors. Further, after removing too long and too short solutions, three candidate Pareto-optimum summaries A, B and C were obtained by the proposed approach.

4.3 Quality of Output Summaries

In this subsection, we study the quality of the obtained summaries A, B and C by comparing them with the summaries for the same document of similar lengths D, E and F created by Mac OSX Summarize, which is an auxiliary function available on a Mackintosh PC. The comparisons are made in terms of the two evaluation measures, that is, the time to comprehend and the remaining information amount. Thus, two male senior students of Aoyama Gakuin University read all the summaries compared for two times and measured the time. Further, they identified which information contents of the original document remain in the summaries.

Table 1 shows the number of characters, the mean time to read, the standard deviation of the time to read, and the number of characters read per second for all summaries compared. It is noticed from the table that the summaries made by the proposed approach can be read faster than those created by Mac OSX Summarize. This means that the proposed approach is capable of providing summaries easier to read and comprehend.

Table 1. Comparisons interms of time

	Proposed system			Mac OSX Summarize		
Summary ID	A	B	C	D	E	F
Number of characters	311	277	263	392	330	230
Mean time to read (s)	33.3	27.9	26.9	43.9	38.0	25.5
Standard deviation of time to read (s)	4.91	5.75	4.01	2.61	4.34	1.26
Number ofcharacters read per second	9.34	9.92	9.78	8.92	8.68	9.02

Fig. 2 represents which information contents remain in each summary. The contents corresponding to the shaded parts in the original document have been judged to remain. It is observed that the summaries created by the proposed approach have covered whole area of the original document. Whereas, those made by Mac OSX Summarize only capture fragmentary information especially when the length is short, since the summarization function takes the extractive method.

Fig. 2. Comparisons in terms of information

Accordingly, it is also confirmed that the quality of the summaries created by the proposed approach measured in terms of the time to read and the remaining information amount is fairly good. However, the judges pointed out that the summaries made by the proposed approach seem like bulleted sentences and do not flow well.

5 Conclusions

This paper proposed a crowdsourcing approach for informative document summarization service. Further, it confirmed that the approach can function properly by a small-scale laboratory experiment using a pilot system implemented as a web application. However, the approach presented in the paper is still a prototype, and has a large room for improvement. For example, promising improvement options include parallelizing the sub-tasks of creation and evaluation, including interaction effects in the objective functions, combining a computerized summarization technique with the crowdsourcing approach, etc. In order to make the sentences in an output summary flow well, in addition to considering interaction effects in the objective functions, introducing another sub-task of adding conjunctions can be effective.

References

1. Afantenos, S., Karkaletsis, V. and Stamatopoulos, P.: Summarization from Medical Documents: A Survey, Artificial Intelligence in Medicine, 33, 157-177 (2005)
2. Gupta, V. and Lehal, G.S.: A Survey of Text Summarization Extractive Techniques, Journal of Emerging Technologies in Web Intelligence, 2, 258-268 (2010)
3. Estelles-Arolas, E. and Gonzalez-Ladron-de-Guevara, F.: Towards an Integrated Crowdsourcing Definition, Journal of Information Science, 38, 189-200 (2012)
4. Law, E. and von Ahn, L.: Human Computation, Synthesis Lectures on Artificial Intelligence and Machine Learning #13, Morgan and Claypool Publishers, (2011)
5. Quinn, A.J. and Bederson, B.B.: Human Computation: A Survey and Taxonomy of a Growing Field, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, 1403-1412 (2011)
6. Crouser, J.R. and Chang, R.: An Affordance-Based Framework for Human Computation and Human-Computer Collaboration, IEEE Transactions on Visualization and Computer Graphics, 18, 2859-2868 (2012)
7. Corney, J.R., Torres-Sanchez, C., Jagadeesan, A.P., Yana, X.T., Regli, W.C. and Medellin, H.: Putting the Crowd to Work in a Knowledge-Based Factory, Advanced Engineering Informatics, 24, 243-250 (2010)
8. Lin, D., Murakami, Y., Ishida, T., Murakami, Y. and Tanaka, M.: Composing Human and Machine Translation Services: Language Grid for Improving Localization Processes, Proceedings of the 7th International Conference on Language Resources and Evaluation, 500-506 (2009)
9. Hu, C., Bederson, B.B. and Resnik, P.: Translation by Iterative Collaboration between Monolingual Users, Proceedings of the ACM SIGKDD Workshop on Human Computation HCOMP 2010, 54-55 (2010)
10. Hu, C., Bederson, B.B., Resnik, P. and Kronrod, Y.: MonoTrans2: A New Human Computation System to Support Monolingual Translation, Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, 1133-1136 (2011)
11. Bernstein, M., Little, G., Miller, R.C., Hartmann, B., Ackerman, M., Karger, D.R., Crowell, D., and Panovich, K.: Soylent: A Word Processor with a Crowd Inside, Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, 313-322 (2010)