Crowdsourcing Strategies for Text Creation Tasks

Ria Mae Borromeo
Keio University
Yokohama, Japan
ria@db.ics.keio.ac.jp

Maha Alsayasneh, Sihem Amer-Yahia,
Vincent Leroy
Univ. Grenoble Alpes, CNRS, LIG,
F-38000 Grenoble, France
firstname.lastname@imag.fr

ABSTRACT
We examine deployment strategies for text translation and text summarization tasks. We formalize a deployment strategy along three dimensions: work structure, workforce organization, and work style. Work structure can be either simultaneous or sequential, workforce organization independent or collaborative, and work style either crowd-only or hybrid. We use Amazon Mechanical Turk to evaluate the cost, latency, and quality of various deployment strategies. We assess our strategies for different scenarios: short/long text, presence/absence of an outline, and popular/unpopular topics. Our findings serve as a basis to automate the deployment of text creation tasks.

Keywords
Crowdsourcing; Text Creation; Deployment Strategies;

1. INTRODUCTION
Crowdsourcing has been applied to all kinds of tasks ranging from the simplest such as image categorization to the most sophisticated such as creating elaborate text. Although several automatic solutions have been designed for text creation, this task remains difficult for machines as it involves a level of abstraction and creativity that only humans are capable of. That is particularly true for translation and summarization where original texts of varying length and complexity need to be understood and processed. In this paper, we examine how hybrid deployment strategies that combine the power of algorithms with the creativity of humans can improve the quality of produced text, as well as the cost and latency of tasks. To the best of our knowledge, our work is the first to explore the effectiveness of hybrid deployment strategies for crowdsourced text creation.

We are interested in two text creation tasks: translation and summarization. It has been shown that for text translation, letting workers edit text and correct each others’ mistakes in a sequential manner, produces higher quality translations than in the case where workers generate independent translations simultaneously [1]. It has also been shown that automatic methods are not very good at summarizing and merging sentences to generate high-quality summaries [11]. We hence propose to study different deployment strategies. A deployment strategy is a plan on how to carry out a task. It is a combination of three dimensions: work structure, workforce organization, and work style. Work structure refers to how a task is deployed among workers, which can either be simultaneous or sequential. Workforce organization refers to how workers are organized to complete a task, which can either be independent or collaborative. Work style distinguishes a hybrid approach, where a task is completed by both algorithms and humans, from a crowd-only approach, where a task is solely carried out by humans. Table 1 shows 6 deployment strategies that combine those dimensions.

The idea of combining humans and machines for task completion has been explored in a variety of domains ranging from databases to machine learning [3, 4, 5, 7, 9, 10, 12]. Our focus is on the evaluation of how our strategies affect cost, latency, and quality of output text. For translation, in addition to work structure, workforce organization, and work style, we pay attention to the properties of the text that is being translated or summarized and study the impact of text length. For summarization, we study the quality of summaries in the presence and absence of a suggested outline, and for topics of varying popularity.

The paper is organized as follows. Our tasks and deployment strategies are given in Section 2. Our experiments are presented in Section 3. We conclude and discuss perspectives raised by this work in Section 4.

2. TASK DEPLOYMENT
2.1 Translation
We examine two types of translation tasks: full document and short text translation. In the first case, the original text is a speech by President Obama entitled “Giving Every Student an Opportunity to Learn Through Computer Science for All.” It consists of 35 sentences and 10 paragraphs. The target language is French. In the second case, the original text is a poem in Arabic, “When You Decide to Leave” by Mahmoud Darwish, with 4 sentences. The target languages are English and French.

2.2 Summarization
We chose movie reviews and soccer games to be summarized into free-text, structured, or personalized summaries. A free-text summary is generic and has no specific struc-
Sequential-Independent-Hybrid (SEQ-IND-HYB) & An initial output is generated automatically then it is sent to one worker at a time for improvement. The final result is a single output.  
Sequential-Independent-CrowdOnly (SEQ-IND-CRO) & An initial output is completed by a worker then it is sent to one worker at a time to improve it. The final result is a single output.  
Simultaneous-Independent-Hybrid (SIM-IND-HYB) & An initial output is generated automatically then sent to several independent workers for improvement. The best output is chosen after an evaluation.  
Simultaneous-Independent-CrowdOnly (SIM-IND-CRO) & Several outputs are created simultaneously by independent workers. The best output is chosen after an evaluation.  
Simultaneous-Collaborative-Hybrid (SIM-COL-HYB) & An initial output is generated automatically then sent to one group of workers who collaborate to improve it.  
Simultaneous-Collaborative-CrowdOnly (SIM-COL-CRO) & One output is created by one group of workers together.  

<table>
<thead>
<tr>
<th>Strategy Description</th>
<th>Strategy</th>
</tr>
</thead>
</table>
| Sequential-Independent-Hybrid | An initial output is generated automatically then it is sent to one worker at a time for improvement. The final result is a single output.  
Sequential-Independent-CrowdOnly | An initial output is completed by a worker then it is sent to one worker at a time to improve it. The final result is a single output.  
Simultaneous-Independent-Hybrid | An initial output is generated automatically then sent to several independent workers for improvement. The best output is chosen after an evaluation.  
Simultaneous-Independent-CrowdOnly | Several outputs are created simultaneously by independent workers. The best output is chosen after an evaluation.  
Simultaneous-Collaborative-Hybrid | An initial output is generated automatically then sent to one group of workers who collaborate to improve it.  
Simultaneous-Collaborative-CrowdOnly | One output is created by one group of workers together.  

Table 1: Deployment Strategies

3. VALIDATION

In this section, we report the setup and the results of experiments we performed to evaluate our proposed deployment strategies. We deployed our tasks on Amazon Mechanical Turk (AMT). The list of required skills was provided at the beginning of each task. In the case of hybrid strategies, we used Google Translate to obtain machine translations and MEAD\(^1\) to obtain automatic summaries. We observed how our strategies affect the cost, latency and result quality. We calculated the cost by taking the sum of all the payments to workers for all the HITs posted to carry out a strategy. We asked experts to rate quality of each text output using a 5-pt Likert scale (1 - very poor, 2 - poor 3 - barely acceptable, 4 - good, 5 - very good) using the following criteria: spelling, syntax, semantic coherence, and adequation to the original text. The latency was derived by adding the amount of time it took for a worker or group of workers to complete each task in a given strategy. Table 2 summarizes the comparisons that we performed.

3.1 Translation

All strategies were considered to translate Obama’s speech. For Darwish’s poem, we only report results for simultaneous work structure and independent workforce organization (Figure 1b). Sequential strategies were not useful since the text is short. Similarly, for collaborative strategies, the time and effort of recruiting workers outweighs their benefit for short text.

Setup. Figure 1a shows how we implemented sequential independent strategies for translation tasks. In the case of a hybrid work style, we first obtained an automatic translation of the original text to the target language. It was then improved by three different workers one after the other. To improve a translation, we published a Human Intelligence Task (HIT) that instructs a worker to enhance an automatically produced translation. For every response, we asked an expert to rate the improved translation. When the rating was good enough, we asked the next worker to enhance the current translation. Otherwise, we asked another worker to enhance the initial translation until we received an acceptable translation. For a crowd-only work style, we first published a HIT that requests a translation of the original text from scratch. After receiving an initial translation, we asked two more workers to improve the translation iteratively.

As shown in Figure 1b, for simultaneous independent strategies, we posted a HIT requesting three workers to translate text simultaneously. For the hybrid work style, workers improved an initial machine translation, while in the crowd-only case, they translated the original text from scratch. After receiving all three answers, we asked an expert to select the best one.

For simultaneous collaborative strategies (Figure 1c), we needed workers to collaborate to create (crowd-only) or improve (hybrid) a translation. We deployed these tasks by

\(^1\)http://www.summarization.com/mead/
post a HIT that explains to workers the task requirements and asks them if they are willing to work on the task with other workers. After that, we invited at least two workers to use Google Docs to collaborate on the translation.

We based our incentives on the pricing scheme in [13] that paid $0.10 (US dollars) per sentence. For the independent tasks, we paid $3.50/HIT for each translation from scratch and $1.75 for each translation improvement HIT. For the collaborative tasks, we paid $1.16 per worker for the translation from scratch HIT and $0.58 per worker for the translation improvement HIT.

Findings. We find that letting workers collaborate as a group has a positive impact on the behavior of workers, which also contributes to raising translation quality. Another advantage of collaboration is a much lower cost, while latency only slightly increases. For translating long text, a hybrid work style combined with a sequential work structure are best (SEQ-IND-HYB). For short text, however, a simultaneous work structure is more appropriate, and both hybrid and crowd-only work styles perform well (SIM-IND-HYB and SIM-IND-CRO).

3.2 Summarization

It has been shown that providing a narrative outline improves text summaries [6]. To verify this finding for movies and soccer games, we crowdsourced summaries with various deployment strategies in the presence and the absence of a proposed summary outline for topics of varying popularity. The summarization tasks were deployed using the same strategies as in Figures 1a, 1b, and 1c with movie reviews or game statistics as input, and a summary text as output. For the hybrid work style, we first obtained an automatic summary using MEAD and gave it to workers for improvements. For the sequential work style, we produced an automatic summary that conform to their expectations (personalized). One can see that ours is generic and covers the main aspects of a movie while that of the worker is more specific. We deployed all strategies for the movie "The Imitation Game" to obtain free-text summaries as well as personalized summaries. The two other movies were summarized with crowd-only work styles, using a structured summary. The incentives we provided for the independent creation of free-text summaries are as follows: $5.00 for each written from scratch, $1.25 for its 1st improvement and $0.62 for the 2nd, $2.50 for the 1st improvement of an automatically generated summary, $0.62 for the 2nd, and $0.31 for the 3rd. For collaborative tasks, we paid each worker $1.16 to create a summary from scratch and $0.58 each to improve a summary. For the structured and personalized summaries of movie reviews, we paid $0.70 for the task of coming up with an outline and for creation and improvement tasks.

Movie Setup. We selected movies with different ratings and popularity. In addition to a structured outline that we provided, we asked three different workers to propose an outline that conforms to their expectations (personalized). Figure 2 shows two example outlines. On the left side is our proposed outline. On the right side is one that a worker suggested. One can see that ours is generic and covers the main aspects of a movie while that of the worker is more specific. We deployed all strategies for the movie "The Imitation Game" to obtain free-text summaries as well as personalized summaries. The two other movies were summarized with crowd-only work styles, using a structured summary. The incentives we provided for the independent creation of free-text summaries are as follows: $5.00 for each written from scratch, $1.25 for its 1st improvement and $0.62 for the 2nd, $2.50 for the 1st improvement of an automatically generated summary, $0.62 for the 2nd, and $0.31 for the 3rd. For collaborative tasks, we paid each worker $1.16 to create a summary from scratch and $0.58 each to improve a summary. For the structured and personalized summaries of movie reviews, we paid $0.70 for the task of coming up with an outline and for creation and improvement tasks.

Findings. We find that workers produce better summaries when given an outline that serves as a template. This finding reinforces previous results in narrative theory that show an increase in emotional worker engagement, and the likelihood of workers sharing those summaries when narrative templates are used to produce them [6]. However, there is a fine line between providing outlines that are general and outlines that ask for specific content requiring workers to spend extra time finding that content. We also observe that a hybrid work style that provides workers with an automatically generated initial summary helps workers structure their thoughts. Among our proposed strategies, we found that SEQ-IND-HYB is best for free-text summaries while SEQ-IND-CRO is best for structured ones. Finally, we find that summarizing reviews for a popular movie does not guarantee a high-quality outcome.

Table 2: Comparison Scenarios

<table>
<thead>
<tr>
<th>Workforce Organization (IND vs COL)</th>
<th>Work Structure (SIM vs SEQ)</th>
<th>Work Style (HYB vs CRO)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIM-IND-HYB vs. SIM-COL-HYB</td>
<td>SIM-IND-HYB vs. SEQ-IND-HYB</td>
<td>SEQ-IND-CRO vs. SEQ-IND-HYB</td>
</tr>
<tr>
<td>SIM-COL-HYB vs. SIM-COL-CRO</td>
<td></td>
<td>SIM-COL-HYB vs. SIM-COL-CRO</td>
</tr>
</tbody>
</table>
4. SUMMARY AND PERSPECTIVES

The main takeaway is that humans have an aversion to long text and to the effort of creating text from scratch (case of full document translation). They are however better than machines at sequentially improving automatically translated text, or at creating text based on outlines (case of summarizing movie reviews). For short text, providing an initial machine translation does not help.

The popularity of an event affects the quality of obtained summaries (case of soccer games). Its recency impacts the resulting quality drops as it becomes necessary to focus on recent results. We requested summaries for two games: one between two popular teams and another between less popular teams. The soccer games were summarized with crowd-only work styles, using a structured summary. To create structured summaries for soccer games, we paid $1.40 and to improve a summary, we also paid $0.70.

Soccer Findings. We observed that the summaries created for popular teams were completed faster and were of higher quality compared to the less popular game. We also noticed that workers prefer working independently and sequentially (SEQ-IND-CRO), as they tend to disagree a lot on this topic, which makes collaborating difficult.

4. SUMMARY AND PERSPECTIVES

The main takeaway is that humans have an aversion to long text and to the effort of creating text from scratch (case of full document translation). They are however better than machines at sequentially improving automatically translated text, or at creating text based on outlines (case of summarizing movie reviews). For short text, providing an initial machine translation does not help.

The popularity of an event affects the quality of obtained summaries (case of soccer games). Its recency impacts the speed at which workers respond. Also, for tasks requiring creativity, and when the input text is short (case of the short poem), humans are best. The same is true when guidelines are provided for text creation tasks (case of summary outlines). However, when those guidelines are too specific, the resulting quality drops as it becomes necessary to focus on finding answers to specific questions (case of personalized summary outlines).

Our findings can serve as a basis for the development of automatic task deployment text creation. In particular, we would like to design an environment that lets requesters interact with suggested deployment strategies and refine them as tasks are completed. This requester-in-the-loop perspective will provide more transparency in crowdsourcing.

5. REFERENCES


