Poster Paper

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 asses 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-

Strategy	Description
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.
Sequential-Collaborative-CrowdOnly (**SIM-COL-CRO**)	One output is created by one group of workers together.

Table 1: Deployment Strategies

ture while the structured and personalized ones are based on a given outline. A structured summary, however, puts more emphasis on the organization of text, while in a personalized summary the content is given primary importance. The choice of movies and soccer allows us to control topic popularity.

For movies, we used IMDb datasets, and chose "The Imitation Game," "2012," and "The Count of Monte Cristo" as they respectively satisfy the following characteristics: popular with high ratings, popular with low ratings, and not popular. For each movie, we selected five reviews with 7 to 10 sentences each, to be summarized in at most 7 sentences.

For soccer, we asked to summarize game statistics into 14 sentences. We chose two games that were recently held at La Liga-Spain 2016: one between two popular teams, Barcelona and Granada, and the other between less popular teams, Rayo Vallecano and Levante.

2.3 Deployment Strategies

Figure 1 illustrates all strategies for the translation task. For instance, SEQ-IND-HYB first generates an initial translation from English to French using Google Translate, then it asks three workers to improve the translation one after the other. In addition to the original text and task instructions, a requester must consider the following: the number of workers to recruit for the task and the result quality requirement, which are affected by time and budget constraints. For example, in translating Obama's speech, a requester may expect the highest possible quality that three workers can achieve within no particular time and without budget restrictions.

Since we want the highest possible quality, we evaluate every response received. The evaluation may be done by experts, by algorithms [8], or by the crowd [2].

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

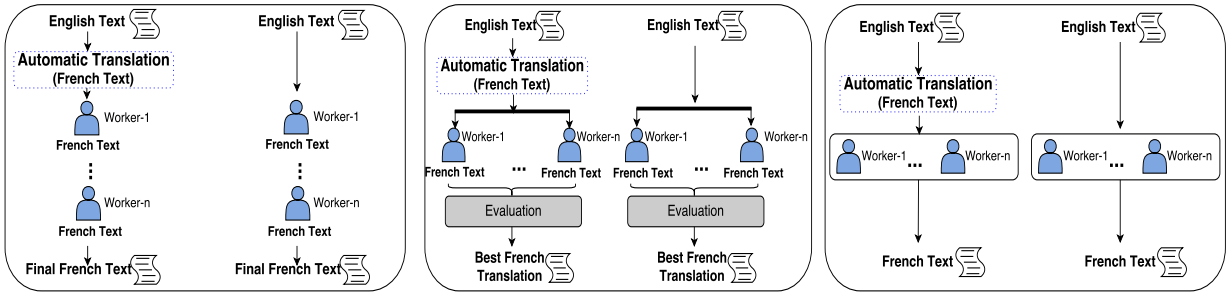
All strategies were considered to translate Obama's speech. For Darwish's poem, we only report results for *simultaneous* work structure and *independent* workforce organization (Figure 1b). Sequential strategies were not useful since the text is short. Similarly, for collaborative strategies, the time 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/*



(a) SEQ-IND-HYB & SEQ-IND-CRO (b) SIM-IND-HYB & SIM-IND-CRO (c) SIM-COL-HYB & SIM-COL-CRO

Figure 1: Translation strategies

Workforce Organization (IND vs COL)	Work Structure (SIM vs SEQ)	Work Style (HYB vs CRO)
SIM-IND-HYB vs. SIM-COL-HYB	SIM-IND-HYB vs. SEQ-IND-HYB	SEQ-IND-CRO vs. SEQ-IND-HYB
SIM-IND-CRO vs. SIM-COL-CRO	SIM-IND-CRO vs. SEQ-IND-CRO	SIM-IND-CRO vs. SIM-IND-HYB
		SIM-COL-HYB vs. SIM-COL-CRO

Table 2: Comparison Scenarios

posting 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. In the case of a *crowd-only* work style, workers were instead provided an outline to follow when producing a summary. This work was then performed with different workforce organizations and work structures, similar to translation 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 1^{st} improvement and \$0.62 for the 2^{nd}; \$2.50 for the 1^{st} improvement of an automatically generated summary, \$0.62 for the 2^{nd}, and \$0.31 for the 3^{rd}. 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 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.

Paragraph 1 – (Introduction) This gives an overview of who is in the film and what it's about. (1 sentence)	**Paragraph 1** – Begin with your overall impression of the movie. (1 sentence)
Paragraph 2 – Describe the plot and the action, while informing the reader which actor plays which role. (3 sentences)	**Paragraph 2** – Mention which actor or plot element impressed you the most. (2 sentences)
Paragraph 3 – Talk about the director and then the actors. Look at good things as well as bad things. (2 sentences)	**Paragraph 3** – Briefly, describe the plot without going into too much detail or including spoilers. (4 sentences)
Paragraph 4 – (Conclusion) Tell the reader whether or not to go and see the film. (1 sentence)	**Paragraph 4** – Conclude by referencing similar movies you enjoyed and encouraging/discouraging people from seeing the movie. (3 sentences)

Figure 2: Two Summarization Outlines

Soccer Setup. We sought to verify how the popularity of a team and the deployment strategy affect the 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

[1] P. André, R. E. Kraut, and A. Kittur. Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 139–148. ACM, 2014.

[2] C. Callison-Burch. Fast, cheap, and creative: evaluating translation quality using amazon's mechanical turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 286–295. Association for Computational Linguistics, 2009.

[3] J. Fan, M. Lu, B. C. Ooi, W.-C. Tan, and M. Zhang. A hybrid machine-crowdsourcing system for matching web tables. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 976–987. IEEE, 2014.

[4] M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, and R. Xin. Crowddb: answering queries with crowdsourcing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 61–72. ACM, 2011.

[5] D. Haas, J. Ansel, L. Gu, and A. Marcus. Argonaut: macrotask crowdsourcing for complex data processing. *Proceedings of the VLDB Endowment*, 8(12):1642–1653, 2015.

[6] J. Kim and A. Monroy-Hernandez. Storia: Summarizing social media content based on narrative theory using crowdsourcing. *arXiv preprint arXiv:1509.03026*, 2015.

[7] G. Li, J. Wang, Y. Zheng, and M. Franklin. Crowdsourced data management: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, PP(99), 2016.

[8] N.-Q. Luong, L. Besacier, and B. Lecouteux. Towards accurate predictors of word quality for machine translation: Lessons learned on french–english and english–spanish systems. *Data & Knowledge Engineering*, 96:32–42, 2015.

[9] A. Marcus, E. Wu, D. R. Karger, S. Madden, and R. C. Miller. Crowdsourced databases: Query processing with people. In *5th Biennial Conference on Innovative Data Systems Research*. CIDR, 2011.

[10] A. G. Parameswaran, H. Park, H. Garcia-Molina, N. Polyzotis, and J. Widom. Deco: declarative crowdsourcing. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1203–1212. ACM, 2012.

[11] G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207, 1997.

[12] H. Wu, H. Sun, Y. Fang, K. Hu, Y. Xie, Y. Song, and X. Liu. Combining machine learning and crowdsourcing for better understanding commodity reviews. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[13] O. F. Zaidan and C. Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1220–1229. Association for Computational Linguistics, 2011.